Journal of Molecular Graphics and Modelling 29 (2010) 518–528

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

# 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

Lotfollah Saghaie[a,b], Mohsen Shahlaei[c,a], Armin Madadkar-Sobhani[d], Afshin Fassihi[a,b,*]

[a] Department of Medicinal Chemistry, Faculty of Pharmacy and Pharmaceutical Sciences, Isfahan University of Medical Sciences, Isfahan, Iran
[b] Isfahan Pharmaceutical Sciences Research Center, 81746-73461 Isfahan, Iran
[c] Department of Medicinal Chemistry, School of Pharmacy, Kermanshah University of Medical Sciences, Kermanshah, Iran
[d] Department of Bioinformatics, Institute of Biophysics and Biochemistry, University of Tehran, Tehran, Iran

ARTICLE INFO

*Article history:*
Received 2 May 2010
Received in revised form
25 September 2010
Accepted 4 October 2010
Available online 12 October 2010

*Keywords:*
Partial least squares
Radial basis function neural networks
Multivariate imaging analysis-quantitative structure activity relationship
Cyclin dependent kinase 4 inhibitors

ABSTRACT

The detailed application of multivariate image analysis (MIA) method for the evaluation of quantitative structure activity relationship (QSAR) of some cyclin dependent kinase 4 inhibitors is demonstrated. MIA is a type of data mining methods that is based on data sets obtained from 2D images. The purpose of this study is to construct a relationship between pixels of images of investigated compounds as independent and their bioactivities as a dependent variable. Partial least square (PLS) and principal components-radial basis function neural networks (PC-RBFNNs) were developed to obtain a statistical explanation of the activity of the molecules. The performance of developed models were tested by several validation methods such as external and internal tests and also criteria recommended by Tropsha and Roy. The resulted PLS model had a high statistical quality ($R^2$ = 0.991 and $R^2_{CV} = 0.993$) for predicting the activity of the compounds. Because of high correlation between values of predicted and experimental activities, MIA-QSAR proved to be a highly predictive approach.

## 1. Introduction

Cyclins, cyclin-dependent kinases (CDKs), and CDK inhibitors (CKIs) all are agents that work simultaneously to regulate the cell cycle phenomena. For sureness about progression of cell cycle phases, formation of complexes between cyclins and CDKs is necessary and CKIs have an inhibitory role against this property [1].

One of the most important problems that occur through different phases of cell cycle is genetic aberrations in the regulatory pathways [2]. The relationship between protein retinoblastoma (pRb) pathway and cancer is determined [3–7]. Furthermore, it seems that irregularities in pRb/p16/cyclin D1/cyclin dependent kinase 4 (Cdk4) pathways occur frequently in human malignant and benign tumors [8]. For this reason, compounds with CDKs inhibitory properties, have high potential to be introduced as anticancer agents which specifically target the cell cycle. It must be noted that a number of CDIs are currently in clinical trials [9–12]. Although a large number of such inhibitors have been recognized and evaluated, there is still a strong attention in developing selective CDIs for their potential power to be utilized as anticancer agents.

The fact that misregulation of CDK4 activity can cause cancer proposes the CDK4 as an important object for intervention in cancer therapy [13].

The use of computational approaches for the estimation of the activity of various molecules as drug candidates prior to their synthesis accelerates drug discovery procedure. The purpose in quantitative structure activity relationship (QSAR) methodology is to construct a relationship between physicochemical properties as independent and bioactivity of ligands as a dependent variable. Physicochemical properties could be obtained as descriptors and then in silico methods are applied to manipulate the information, remove noise and derive useful information. Most descriptors may be meaningful physicochemically, so interpretation of the developed models is possible whilst some others do not have direct physicochemical meaning, but contain useful information. Principal components belong to the latter type of descriptors and could be treated and used in QSAR.

Comparative field analysis (CoMFA) is one of the branches of 3D QSAR. In this method understanding the major conformation(s) and position(s) of ligands in the vicinity of active site is necessary. This information helps the computational chemist produce

[*] Corresponding author at: Department of Medicinal Chemistry, Faculty of Pharmacy, Isfahan University of Medical Sciences, 81746-73461 Isfahan, Iran. Tel.: +98 311 7922562; fax: +98 311 6680011.
*E-mail address:* fassihi@pharm.mui.ac.ir (A. Fassihi).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.10.001

descriptors appropriate to signify molecular interaction energies between the ligand and receptor. Some of these energies are steric and electrostatic fields surrounding the molecules. Hydrogen bonds between ligand(s) and active site are also determined by this type of information. Statistics is then computed by multivariate calibration method(s) and the output is displayed as contours superimposed on the molecules.

This method is comprehensive and less trustworthy or even impracticable for big molecules with a lot of bonds. Hence methods that can be replaced with CoMFA to avoid these types of complication are very important. Multivariate image analysis applied in quantitative structure activity relationship (MIA-QSAR) is one of the best suggestions because neither specific tools nor high computations are needed.

Multivariate image analysis is a type of multivariate regression methods that is based on data sets obtained from 2D images. In multivariate image analysis applied in QSAR (MIA-QSAR) images of bioactive molecules are generated and used.

Images can be considered as rich sources of information that have a wide variety of applications in different branches of science including chemistry, remote sensing, geology, agriculture and quality control in food and pharmaceutical production [14]. Images can be divided to univariate and multivariate types. In univariate or gray-scale images, each image is a 2D-matrix with height × width dimensions. In multivariate images each image is a 3D array with height × width × wavelength dimensions. The most common type of multivariate images is color image in which wavelengths corresponding to red, green and blue lights are measured respectively, therefore dimensions of 3D array of these types of images are height × width × 3. After building binaries of each image, they are superimposed to generate a tensor. The generated tensor is unfolded in order to use two-way analysis methods. This generated 2D matrix can be analyzed in the same ways as usual multivariate data sets and data mining methods could be used to obtain useful information among images of molecules.

It is worthy of note that data mining approaches range from simple parametric models derived from linear techniques to complex non-linear models derived from non-linear techniques.

Another important subject in image analysis is the similarities between the molecular structures used. Since the descriptors used in model building are pixels of the images of the molecules, problem of collinear descriptors is very serious in MIA-QSAR. Thus methods on the basis of orthogonalization of original variables such as PCA and PLS could be used. Theses types of regression methods that can use collinear descriptors are very useful in analysis of data sets applied in MIA-QSAR.

Artificial neural networks (ANNs), as a non-linear regression approach, has been widely used to investigate and solve various problems in various branches of science including QSAR [15,16]. ANNs can learn the complex relationships between given inputs and outputs, serving as a useful tool for constructing regression and classification, signal processing, and optimization models.

The main advantage of artificial neural networks over other regression methods is that ANNs act as a black box model, that is, ANNs do not need any governing equation with forceful assumptions specifically describing the underlying data set.

Presently the most widely used network type in ANN-based QSAR reports is neural networks trained by back-propagation (BP) learning algorithm. This algorithm has some disadvantages such as being caught in local minima during learning, very poor convergence rate, time-consuming procedures, and difficulty in explicit optimum network configuration [17].

The radial basis function neural networks (RBFNNs) let us construct regression model between PCs and activity using a fast linear approach. RBFNNs have advantages of short training time and reaching to the optimal unique solution by attaining the global minimum of error surface during training of network. The topology and parameters of developed RBFNNs are straightforward to optimize [18,19].

We decided to use PLS as one of the commonly used linear approaches as well as neural networks as one of the most common non linear approaches for data mining. We used PCs for nonlinear model building to avoid the collinearity problem and modeling of PCs with RBFNNs was referred as PC-RBFNNs.

In the present investigation, we are going to report the combination of multivariate image analysis and RBFNNs. To the best of our knowledge there is not any report for such a combination of statistical methods in QSAR and QSPR studies. These methods were applied for the computation of activity of indenopyrazole derivatives using pixels of bitmaps of molecules as input for QSAR model building. The prediction results were very satisfactory in both training and test set compounds.

## 2. Experimental

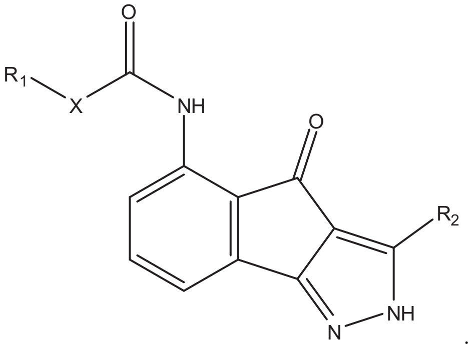
### 2.1. Descriptor generation and assigning training and test sets

In vitro biological activity data used in this study were CDK4 inhibitory activity (in terms of-log $IC_{50}$), of a set of ninety four indenopyrazole derivatives selected from literature [20–24]. General chemical structures and the structural details of these compounds are given in Table 1.

A Pentium IV personal computer (CPU at 2.6 GHz) with Windows XP operating system was used. The two dimensional structures of ninety four molecules were built using ChemDraw 7.0 (ChemDraw Ultra, 1985–2001; CambridgeSoft, Cambridge, MA, USA), and then saved in bitmaps. Afterwards the bitmap of molecules were set to 940 × 600 pixels windows, with resolution of 96 × 96 points per inch. Since the bitmaps of molecules should be superimposed as a 2D alignment, a common pixel was selected among the whole series of molecular structures and then the molecules were totally fixed in that given coordinate. The given coordinate used in the alignment procedure of molecules is shown in Fig. 1. After transferring the bitmaps to the Matlab (version 7.5, 2007; MathWorks, Natick, MA) environment, each 2D image was converted into binaries, a double array in Matlab. The 3D array, the predictors block, was built by grouping the 94 images, giving a 94 × 940 × 600 array. The 3D array was unfolded to a two way array (80 × 564,000). This array was applied in order to be correlated with the dependent variable that was the vector of activities, $pIC_{50}$ of studied molecules. Partial least-squares (PLS) regression as a linear method and radial basis function neural networks as a nonlinear method were applied and the results were compared. Before performing the regressions, in order to minimize the memory used, columns with zero variance in predictor matrix were deleted to generate final $X$ matrix with 94 rows and 14,775 columns. Hence the calculated descriptors were arranged in an $X$ matrix so that the number of rows and columns were the number of molecules and non zero variance pixels, respectively.

After building of $X$ matrix about 20% of the molecules (20 out of 94) were selected as test set molecules for the evaluation of performance of generated regression methods. The best way of assigning test and training sets is dividing data set to guarantee that both sets individually cover the total space occupied by original data set. Ideal splitting of data set is performed in such a way that each of the objects in test set is 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 into the training and test sets. According to Tropsha et al. the best models would be built when Kennard and Stone algorithm is used [25]. Thus, this algorithm was applied in this study [26]. Since we used two various regression methods, and input vectors of the two regression methods were different (latent variables for PLS and principal components for RBFNNs), different

**Table 1**
Structures and details of the molecules used in this study.



| Compd. | $R_1$ | X | $R_3$ |
|---|---|---|---|
| 1 | $(CH_3)_2$ | CH | –Ph-4-OMe |
| 2 | $(CH_3)_3$ | C | –Ph-4-OMe |
| 3 | $(Me)_2N$- | $CH_2$ | –Ph-4-OMe |
| 4[a,b] | Morpholine-4-yl | $CH_2$ | –Ph-4-OMe |
| 5 | Piperazin-1-yl | $CH_2$ | –Ph-4-OMe |
| 6 | Ethyl-NH | $CH_2$ | –Ph-4-OMe |
| 7[b] | N-methyl piperazine | $CH_2$ | –Ph-4-OMe |
| 8 | 4-Aminomethylpiperidine | $CH_2$ | –Ph-4-OMe |
| 9[a] | 4-Amidopiperidine | $CH_2$ | –Ph-4-OMe |
| 10 | 4-Hydroxylmethylpiperidine | $CH_2$ | –Ph-4-OMe |
| 11[b] | 4-Amidopiperazine | $CH_2$ | –Ph-4-OMe |
| 12[a] | 4-Amidinopiperazine | $CH_2$ | –Ph-4-OMe |
| 13[b] | H | $CH_2$ | –Ph-4-OMe |
| 14 | Benzyl | NH | –Ph-4-OMe |
| 15 | Phenyl | NH | –Ph-4-OMe |
| 16 | n-Butyl | NH | –Ph-4-OMe |
| 17[b] | $(Me)_2NNH$ | NH | –Ph-4-OMe |
| 18[a] | 4-Methylpiperazine | NH | –Ph-4-OMe |
| 19[b] | Morpholine-4-yl | NH | –Ph-4-OMe |
| 20[b] | Piperidin-1-yl | NH | –Ph-4-OMe |
| 21 | Pyrrolidine-1-yl | NH | –Ph-4-OMe |
| 22 | H | $CH_2$ | –Ph-4-OMe |
| 23 | H | $CH_2$ | –Ph-4-OMe |
| 24 | H | $CH_2$ | –Ph-4-Et |
| 25 | H | $CH_2$ | –Ph-4-n-Pr |
| 26[b] | H | $CH_2$ | –Ph-4-OH |
| 27 | –Ph-4-$NH_2$ | $CH_2$ | –Ph-4-OMe |
| 28 | H | $CH_2$ | –Ph-4-$NMe_2$ |
| 29 | H | $CH_2$ | –Ph-4piperidino |
| 30 | H | $CH_2$ | –Ph-4-morpholino |
| 31[a] | H | $CH_2$ | –Ph-4-SMe |
| 32[a,b] | Morpholino | $CH_2$ | –Ph-4-$NMe_2$ |
| 33 | 4-(OH)piperidine-1-yl | $CH_2$ | –Ph-4-$NMe_2$ |
| 34 | 4-(Aminomethyl) piperidin-1-yl | $CH_2$ | –Ph-4-$NMe_2$ |
| 35[b] | N-Methylpiperazin-1-yl | $CH_2$ | –Ph-4-$NMe_2$ |
| 36 | Morpholino | $CH_2$ | –Ph-4-morpholino |
| 37 | 4-(OH) piperidine-1-yl | $CH_2$ | –Ph-4-morpholino |
| 38 | 4-(Aminomethyl) piperidin-1-yl | $CH_2$ | –Ph-4-morpholino |
| 39 | H | NH | 3-thienyl |
| 40[a,b] | N-Methylpiperazin-1-yl | $CH_2$ | –Ph-4-morpholino |
| 41 | 4-(Aminomethyl) piperidin-1-yl | $CH_2$ | Et |
| 42 | 4-(Aminomethyl) piperidin-1-yl | $CH_2$ | Cyclopropyl |
| 43 | 4-(Aminomethyl) piperidin-1-yl | $CH_2$ | Cyclohexyl |
| 44 | H | NH | Cyclopropyl |
| 45 | H | $CH_2$ | 4-Pyridyl |
| 46 | H | $CH_2$ | 2-Thienyl |
| 47[a,b] | H | NH | 2-Thienyl |
| 48 | H | NH | 2-Thienyl,3-OMe |
| 49 | H | NH | 2-Thienyl,5-Me |
| 50[a,b] | H | NH | 2-Furanyl |
| 51 | H | NH | 2-Thienyl,5-$CO_2Et$ |
| 52 | H | NH | 3-Thienyl,5-Cl |
| 53 | H | NH | 3-Pyrrolyl,1-Me |
| 54[b] | Dimethylamino | NH | 2-Thienyl |
| 55 | Dimethylamino | NH | 5-(OMe) thien-2-yl |
| 56 | Dimethylamino | NH | 5-(Me) thien-2-yl |
| 57 | Dimethylamino | NH | 5-($CO_2Et$) thien-2-yl |
| 58[a,b] | Dimethylamino | NH | 3-Thienyl |
| 59[a] | Dimethylamino | NH | 5-(Cl) thien-3-yl |
| 60 | Dimethylamino | NH | 2,5-(Di-Me) thien-3-yl |
| 61[a] | Dimethylamino | NH | Furan-2-yl |
| 62 | Dimethylamino | NH | 2,4-(Di-Me)thiazol-5-yl |
| 63[a] | Morpholine-4-yl | NH | 5-(Me) thien-2-yl |
| 64 | Morpholine-4-yl | NH | 5-($CO_2Et$) thien-2-yl |
| 65 | Morpholine-4-yl | NH | 5-(Cl) thien-3-yl |
| 66 | 4-(Methyl)piperazin-1-yl | NH | 5-($CO_2Et$) thien-2-yl |
| 67 | 4-(Aminomethyl) piperidin-1-yl | NH | Isopropyl |
| 68 | 4-(Methyl)piperazin-1-yl | NH | 2,5-(Di-Me) thien-3-yl |
| 69 | 4-(Methyl)piperazin-1-yl | NH | 2,4-(Di-Me)thiazol-5-yl |
| 70 | $(Me)_2CHCONH$– | NH | –Ph-4-OMe |
| 71[a] | 4-(OH)Ph$(CH_2)_2$CONH– | NH | –Ph-4-OMe |
| 72 | 4-(OMe)PhCONH– | NH | –Ph-4-OMe |
| 73 | 3-($NO_2$)PhCONH– | NH | –Ph-4-OMe |
| 74 | 3,4,5-(Tri-OMe)PhCONH– | NH | –Ph-4-OMe |
| 75 | 3-(Me)PhCONH– | NH | –Ph-4-OMe |
| 76 | 3,4-(Di-OMe)PhCONH– | NH | –Ph-4-OMe |
| 77 | (4-OH,3-$NH_2$) PhCONH– | NH | –Ph-4-OMe |
| 78[a,b] | 2,5-(Di-Cl)PhCONH– | NH | –Ph-4-OMe |
| 79 | 3,4-(Di-OH)PhCONH– | NH | –Ph-4-OMe |
| 80 | 3,5-(Di-$NH_2$)PhCONH– | NH | –Ph-4-OMe |
| 81[a] | MeOCONH– | NH | –Ph-4-OMe |
| 82 | 2-(OH)PhCONH– | NH | –Ph-4-OMe |
| 83 | Naphthalen-2-yl CONH– | NH | –Ph-4-OMe |
| 84[a,b] | BnCONH– | NH | –Ph-4-OMe |
| 85 | PhCONH– | NH | –Ph-4-OMe |
| 86[a] | 4-PyrridylCONH– | NH | –Ph-4-OMe |
| 87 | 3-PyrridylCONH– | NH | –Ph-4-OMe |
| 88[a,b] | MeCONH– | NH | –Ph-4-OMe |
| 89 | 4-(OH)PHCONH– | NH | –Ph-4-OMe |
| 90[a,b] | $H_2NCOCONH$– | NH | –Ph-4-OMe |
| 91[b] | 3-($NH_2$)PhCONH– | NH | –Ph-4-OMe |
| 92 | 2,4-(Di-OH)PhCONH– | NH | –Ph-4-OMe |
| 93 | 4-$NH_2$PhCONH– | NH | –Ph-4-OMe |
| 94 | H | NH | –Ph-4-OMe |

[a] Molecules were used as external prediction set for PLS by Kennard–Stone algorithm.

[b] Molecules were used as external prediction set for GA-PC-GRNN by Kennard–Stone algorithm.

sets of molecules were assigned by Kennard and Stone algorithm as training and test sets that are shown in Table 1.

### 2.2. Model development

Because of the similarities between the molecular structures used in this study, and since the descriptors used in model building are pixels of the images of the molecules, problem of collinear and noisy descriptors is very serious. Methods such as PCA and PLS regression can use collinear descriptors. These methods generate new orthogonal descriptors, principal components (in PCA) and latent variables (in PLS), probably resulting better predictive models.

For building QSAR models two different regression methods were used: (1) partial least squares and (2) radial basis function neural networks with principal components as input of network.

PLS is a generalization form of multiple linear regression, which can handle regressions problems when data matrix variables are severely correlated and/or include numerous variables [27]. This approach reduces the number of solutions of a regression problem, which is statistically more robust than multiple linear regression. PLS model generates new orthogonal variables (latent variables) which are linear combinations of the original descriptors with weights.

The important step in PLS analysis is the proper unfolding of original 3D matrix and producing a 2D array of molecules. In this new generated matrix, pixels of each molecule were rearranged in rows. Performing PLS, the 2D data matrix was decomposed into score and loading matrix. After using score matrix instead of the original data matrix as predictor, model building was performed. Preprocessing of data matrix such as mean centering and autoscaling was applied.

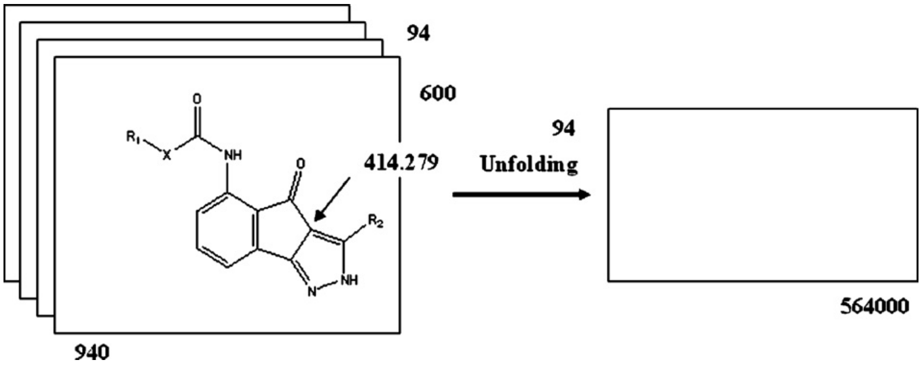


**Fig. 1.** Building of 3D array and 2D array of molecules in this study. The pixel shown by arrow, common to all molecules, was fixed at 414,279 coordinate. Bitmap of molecules was fixed in 940 × 600 pixels windows. The 3D array of molecules was unfolded to 94 × 564000.

The number of latent variables (LVs) used in the partial least squares method is very critical. Using too few LVs in model development will generate an underfitted model, i.e., $X$ block fits with the activity of molecules with a low $R^2$. Using too many of LVs, on the other hand, produces an overfitted model. An overfitted model fits with some of the noise information of the training set thus generates a low error in the prediction of activity but performs poorly in the prediction of activity of molecules in the test set. The optimum number of LVs will then divide the $X$ matrix into the useful information and the noise. Thus a major decision in developing successive QSAR is when to stop adding parameters to the model (here latent variables) during the regression procedure. Selection of optimum number of LVs was done here using the root mean square error of calibration and root mean square error of leave-one-out cross-validation.

In the case of each number of latent variables used, given $n$ molecules in training set, model was built and *RMSEC* was calculated for a given number of LVs. Also for cross validation $n - 1$ molecules were used in the model building step and the resulted model was used to predict the $\text{pIC}_{50}$ of molecule that was not used in model building. This procedure was repeated $n$ times until $\text{pIC}_{50}$ of all molecules were predicted. The *RMSECV* for each number of latent variables was calculated by comparing the predicted $\text{pIC}_{50}$ of molecules with known $\text{pIC}_{50}$ of molecules in training set. Same procedure was used for PC-RBFNNs.

The combination of principal components of $X$ matrix generated from images of molecules as input of radial basis function neural networks is referred as PC-RBFNNs.

In principal component radial basis function neural networks, the original data matrix is reduced to an orthogonal principal component and their scores are used as inputs for RBFNNs. In PC-RBFNNs model, the use of scores instead of original data reduces input nodes, so training time of the network is shortened. Also, noisy information and random error in the original data will be excluded. So using PCs generates a more accurate RBFNNs model.

In the multivariate imaging analysis number of independent variables is very large, so data reduction is very necessary.

The main advantage of radial basis function neural network is that it does not require iterative learning. It is an interesting property that makes the method very attractive for model building and hence RBFNNs is very faster than the well-known back-propagation neural network.

The complete explanation behind the theory of radial basis function neural networks is adequately described elsewhere [28,29].

Here only a brief description of this type of neural network is presented. RBFNNs include three layers: input layer, hidden layer and output layer as presented schematically in Fig. 2. Each neuron in each given layer is fully connected to the next layer but there is not any connection between neuron in a given layer. No processing occurs on the input information in the input layer and the duty of this layer is only distribution of input to the hidden layer. In the hidden layer of RBFNNs there are a number of radial basis function units ($n_h$) and bias ($b_k$). Each hidden layer unit represents a single radial basis function, with associated center position and width. In the hidden layer each neuron applies a radial basis function as nonlinear transfer function to operate on the input information coming from the previous layer. The most often used RBF is Gaussian function that is characterized by a center ($c_j$) and width ($r_j$). By measuring the Euclidean distance between input vector ($x$) and the radial basis function center ($c_j$) the RBF function performs the nonlinear transformation using following equation in the hidden layer:

$$h_j(x) = \exp\left(-\frac{||x - c_j||^2}{r_j^2}\right) \quad (1)$$

where $h_j$ is the notation for the output of the $j$th RBF unit. For the $j$th RBF, $c_j$ and $r_j$ are the center and width, respectively. The operation of the output layer is linear, which is given in:

$$y_k(x) = \sum_{j=1}^{n_h} w_{kj} h_j(x) + b_k \quad (2)$$

where $y_k$ is the $k$th output unit for the input vector $x$, $w_{kj}$ is the weight connection between the $k$th output unit and the $j$th hidden layer unit, and $b_k$ is the bias.

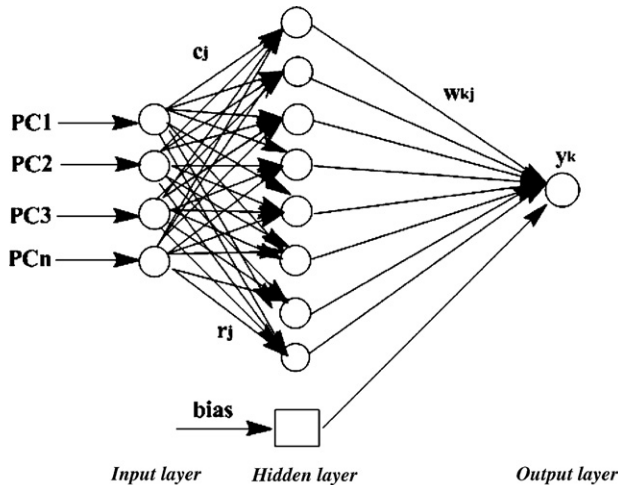


**Fig. 2.** The architecture of PC-RBFNN.

In order to optimize RBFNNs, centers, number of hidden layer units, width, and weights should be selected. Random subset selection, K-means clustering, orthogonal least-squares learning algorithm, and RBF-PLS are various ways for choosing the centers. The same widths of the radial basis function networks for all the units or different widths for each unit could be selected for optimizing RBFNNs.

In this paper, Gaussian functions with a constant width, which was the same for all units were selected. Using training set molecules the centers were optimized by forward subset selection routine. After the selection of optimum values of centers and width of radial basis functions the connection weight between hidden layer and output layer was adjusted using a least-squares solution.

The overall performance of RBFNNs is evaluated in terms of root mean square error cross validation (*RMSECV*) according to the following equation:

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n_s}(y_k - \hat{y}_k)^2}{n_s}} \tag{3}$$

where $y_k$ is the experimental value of biological activity, $\hat{y}_k$ is the output predicted activity of network calculated by cross validation. $n_s$ is the number of compounds in the analyzed set.

### 2.3. Model validation, predictability and robustness of model

To demonstrate that the resulted models have good prediction of activity of selected studied compounds, some different methods of evaluation of model performance have been used. Model performance can be evaluated by different approaches. Here, $R^2$, which presents the explained variance for given set, was used to determine the goodness of model's fit performance. In addition, the prediction performance of the built models must be estimated in order to build a successful QSAR model. In this investigation, we evaluated the prediction performance of developed models using two parameters, the root mean square error (*RMSE*) and predicted error sum of square (*PRESS*(%)).

In order to assess the predictive ability and to check the statistical significance of the developed models, the proposed models were applied for the prediction of values of $\text{pIC}_{50}$ for external set that were not used in model building.

Cross-validation is a technique used to explore the reliability of statistical models. Root mean square error cross validation (*RMSECV*) as a standard index to measure the accuracy of a modeling method which is based on the cross-validation technique and $R^2_{LOO}$ as another criterion of predictability of developed models were applied.

According to Tropsha high $R^2_{LOO}$ does not routinely mean a high predictability of the developed model. Thus, the high value of $R^2_{LOO}$ is the necessary but not the sufficient condition for the developed model to have a high predictability. We reason that in addition to a high $R^2_{LOO}$ a reliable model should also be characterized by a high $R^2$ between the calculated and experimental values of compounds from a test set [30].

Some criteria are suggested by Tropsha. If these criteria were satisfied then it can be said that the model is predictive [25]. These criteria include:

$$R^2_{LOO} > 0.5 \tag{4}$$

$$R^2 > 0.6 \tag{5}$$

$$\frac{R^2 - R_0^2}{R^2} < 0.1; \qquad \frac{R^2 - R'^2_0}{R^2} < 0.1 \tag{6}$$

$$0.85 < k < 1.15 \quad or \quad 0.85 < k' < 1.15 \tag{7}$$

$R^2$ is the correlation coefficient of regression between the predicted and observed activities of compounds in training and test sets. $R_0^2$ is the correlation coefficient for regressions between predicted versus observed activities through the origin, $R'^2_0$ is the correlation coefficient 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 $R_0^2$, $R'^2_0$, $k$ and $k'$ are presented obviously in literature [25].

In addition, according to Roy and Roy [31] the difference between values of $R_0^2$ and $R'^2_0$ must be studied and given importance. They suggested following modified $R^2$ form:

$$R_m^2 = R^2(1 - |\sqrt{R^2 - R_0^2}|) \tag{8}$$

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

In order to avoid chance correlations which are possible because of a large number of generated columns (independent variables), and examine the robustness of developed models, *Y*-randomization test has been applied to models. The dependent variable vector is randomly permuted and a new QSAR models is constructed using the original independent variable matrix. The new modeling was expected to have low $R^2$ values. For sureness, some iterations were carried out. If the results show high $R^2$, it implies that an acceptable QSAR model cannot be obtained.

The performance of our developed PC-RBFNN model was evaluated by measuring the sensitivity (SE), specificity (SP), accuracy (ACC), and Matthew's correlation coefficient (MCC). During the training of nonlinear model, a value of 1 was assigned for the molecules with $\text{pIC}_{50}$ higher than 6.5 (64 molecule in whole data set) and 0 for compounds with $\text{pIC}_{50}$ lower than 6.2 (19 compounds). To evaluate the above parameters for the developed nonlinear models, this model was run for considered molecules and their activities were calculated using 0.1 and 0.9 cutoff values for effective and non effective compounds, respectively. Thus, an effective compound was correctly predicted by PC-RBFNNs when its output value ranged from 0.9 to 1. For each non effective compound (18 compounds) correct prediction of the PC-RBFNNs provided output values between 0.0 and 0.1. If predicted $\text{pIC}_{50}$ for molecules lay in the 0.2–0.9, these predictions were considered as incorrect predictions. The SE, SP, ACC and MCC parameters were calculated using following equations:

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{10}$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{MCC} = \frac{\text{TPTN} - \text{FPFN}}{\sqrt{(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TP} + \text{FP})(\text{TN} + \text{FN})}} \tag{12}$$

where TP stands for true positives (inhibitor compounds predicted as inhibitor); FN denotes false negatives (inhibitor compounds predicted as non inhibitor); TN means true negatives (non inhibitor compounds predicted as non inhibitor) and FP refers to false positives (non inhibitor compounds predicted as inhibitor).

## 3. Results and discussion

As it was discussed in the following sections, generated descriptors in multivariate image analysis do not have any direct interpretation physicochemically. But it should be noted that *X* block generated using these types of descriptors includes useful information that can be treated and used for model building

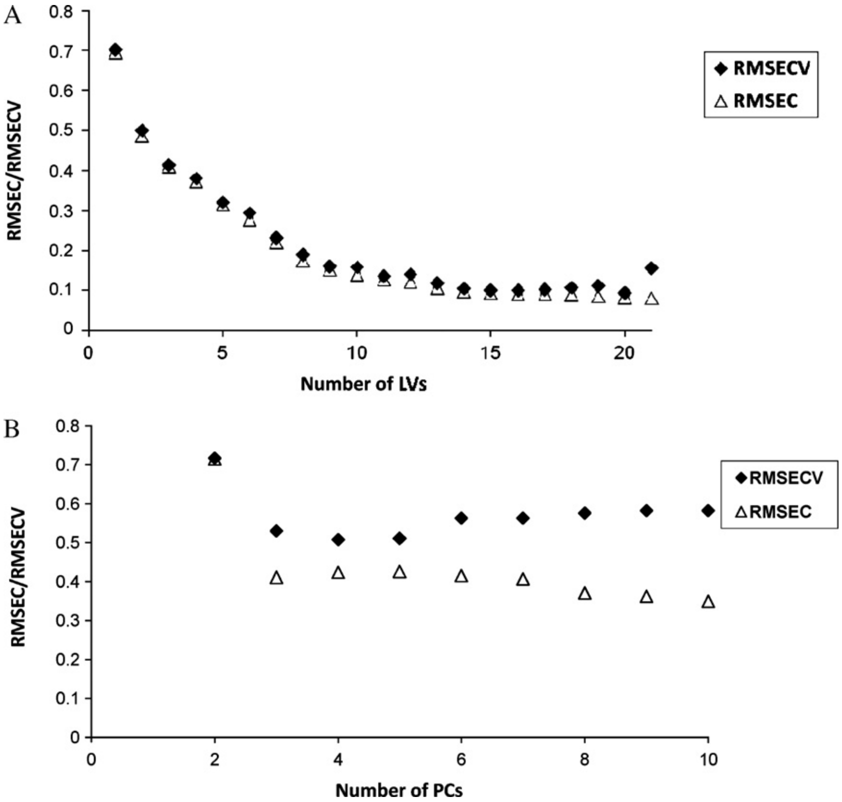


**Fig. 3.** Plots of *RMSECV* and *RMSEC* vs. (A) number of latent variables by PLS and (B) number of principle components by RBFNN.

and prediction of activity of new sets of molecules such as test set.

As discussed in Section 1, MIA-QSAR performs as a good alternative for CoMFA method. Compared with CoMFA, MIA-QSAR uses bitmaps of 2D images of molecules as descriptors. This information does not stand for particular 3D information or interactions such as steric interactions or any other physicochemical descriptors. But, imaging is a general way to signify many properties since very small changes in drawing a given structure can cause significant changes in the predicted property, i.e. each pixel place acts as a code (the descriptors are binary). For instance, R and S enantiomers of a given chiral molecule may be schematically differentiated by drawing wedge or hashed bonds at the chiral center, and this will cause, if is the case, substantial differences in the studied property, for example, biological activity. This means imaging is extremely sensitive to the drawing. In this way, neither simulation of 3D interactions nor odd descriptors, which are difficult to interpret, are necessary in the MIA-QSAR model building.

The bitmaps of 94 molecules were matricized and a lot of descriptors (columns of *X* block) were calculated for each molecule using bitmaps of molecules. Logarithms of the inverse of biological activity ($\log 1/\mathrm{IC}_{50}$) data of 94 molecules were used to get the relationship with independent variables.

Before applying the PLS and RBFNNs, and due to the quality of data, a pretreatment of the original data was necessary. Thus, autoscaling and deletion of columns with zero variance were performed.

After dividing the molecules into two parts, calibration and validation sets, based on Kennard and Stone algorithm, building of PLS and PC-RBFNN models using training set was performed. Developed models were used to predict the activity of molecules in test set to evaluate the performance of the developed models.

### 3.1. PLS analysis

To solve the problem of collinearity in the generated descriptors partial least squares regression as a linear method was used to model structure activity relationship quantitatively.

Two quantities, root mean square error of calibration (*RMSEC*) and root mean square error of cross validation *(RMSECV)* were used for the optimization of a number of the latent variables used in model development. As it is shown in Fig. 3A, 16 latent variables were selected as optimum number of latent variables.

The PLS equation was trained using training data set and it was evaluated by prediction molecules. The predicted activity of molecules calculated by PLS is plotted against the experimental values in Fig. 4A and is reported in Table 2. As was expected, the calculated values are in good agreement with experimental values. The residuals of the PLS predicted values are plotted against the experimental values in Fig. 5A which shows no systematic error in the developed model.

As a result, it was found that correctly designed partial least squares regression method could practically represent dependence of the activity of cyclin dependent kinase 4 inhibitors on the extracted descriptors from bitmaps of molecules.

In Table 3 results of various statistics criteria and figures of merit for this model are reported for two subsets of molecules, i.e. training and test set of molecules. As it is shown in Table 3, $R^2$ that is a criterion of goodness of fit of the proposed model was obtained for two sets. The high value of this parameter indicates a good fit between experimental and predicted values of antagonist activities of compounds by the developed model indicating high capability of the proposed model. As can be seen from this table, the pixels used in this model can explain 99.10% of the variances in the CDK4 inhibitory activity of the compounds used in training set. The external predictability and model capability of a proposed model

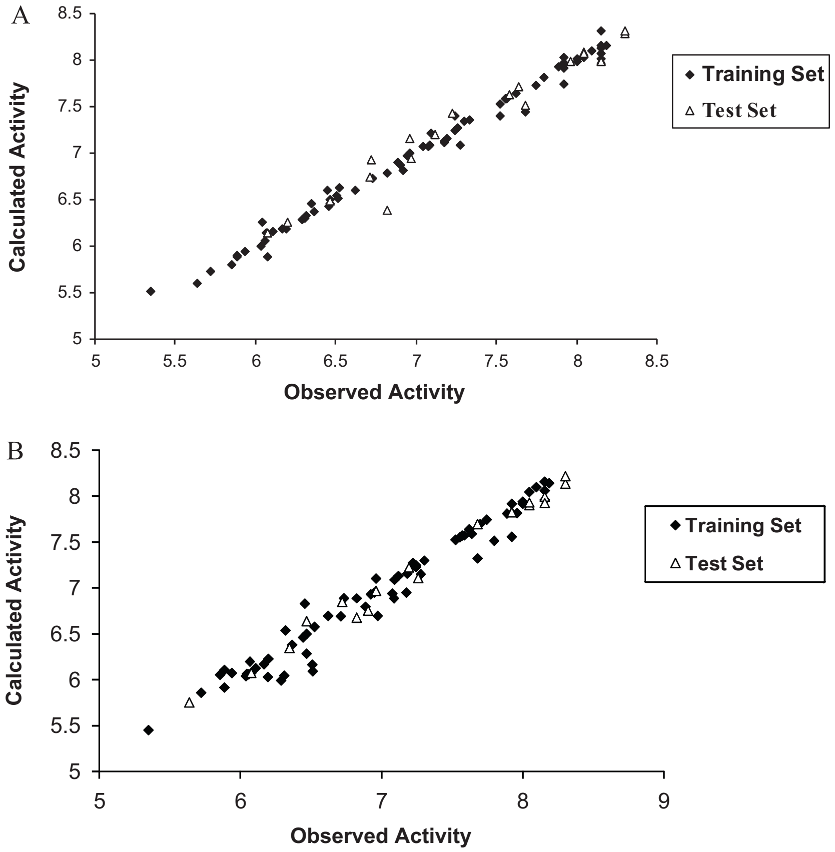


**Fig. 4.** $pIC_{50}$ estimated by various modeling versus experimental values for training and test sets: (A) PLS and (B) PC-RBFNN.

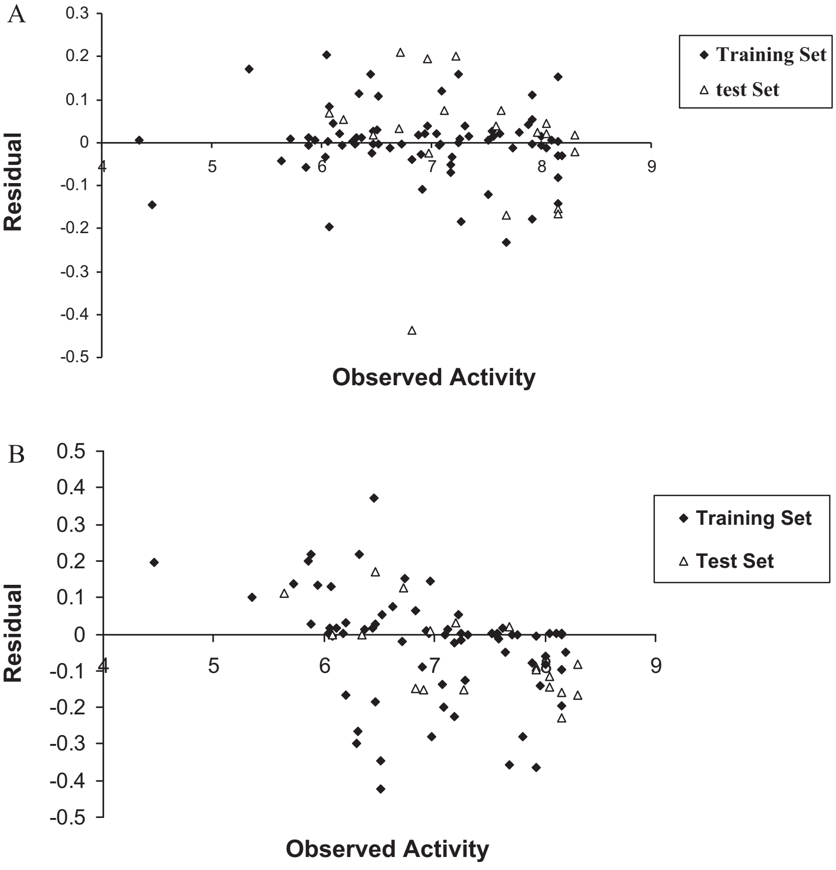


**Fig. 5.** Plots of residuals versus experimental factors: (A) PLS and (B) PC-RBFNN.

**Table 2**
The experimental $pIC_{50}$ and the predicted values of the training set and test set and values of relative error prediction by each model.

| Compound no. | $pIC_{50}$ observed | $pIC_{50}$ predicted (PLS) | REP | $pIC_{50}$ predicted (PC-RBFNN) | REP |
|---|---|---|---|---|---|
| 1 | 4.456 | 4.310 | −0.146 | 5.053 | 0.118 |
| 2 | 4.347 | 4.352 | 0.005 | 5.043 | 0.138 |
| 3 | 6.046 | 6.251 | 0.205 | 6.063 | 0.003 |
| 4 | 6.710 | 6.741 | 0.031 | 6.690 | −0.003 |
| 5 | 6.947 | 6.968 | 0.021 | 6.949 | 0.000 |
| 6 | 6.038 | 6.004 | −0.034 | 6.039 | 0.000 |
| 7 | 6.903 | 6.874 | −0.029 | 6.751 | −0.023 |
| 8 | 7.699 | 7.656 | −0.006 | 7.698 | 0.000 |
| 9 | 7.119 | 7.194 | 0.075 | 7.433 | 0.042 |
| 10 | 7.086 | 7.083 | −0.003 | 6.885 | −0.029 |
| 11 | 7.194 | 7.160 | −0.034 | 7.424 | 0.031 |
| 12 | 7.585 | 7.623 | 0.038 | 7.573 | −0.002 |
| 13 | 6.347 | 6.462 | 0.115 | 6.045 | −0.050 |
| 14 | 6.512 | 6.509 | −0.003 | 6.089 | −0.069 |
| 15 | 6.057 | 6.060 | 0.003 | 6.057 | 0.000 |
| 16 | 6.108 | 6.153 | 0.045 | 6.126 | 0.003 |
| 17 | 7.678 | 7.445 | −0.233 | 6.921 | −0.109 |
| 18 | 8.046 | 8.091 | 0.045 | 7.702 | −0.045 |
| 19 | 7.921 | 7.973 | 0.052 | 7.432 | −0.066 |
| 20 | 7.921 | 8.033 | 0.112 | 7.426 | −0.067 |
| 21 | 7.921 | 7.918 | −0.003 | 7.156 | −0.107 |
| 22 | 6.076 | 5.880 | −0.196 | 6.071 | −0.001 |
| 23 | 6.310 | 6.306 | −0.004 | 6.044 | −0.044 |
| 24 | 6.194 | 6.187 | −0.007 | 6.026 | −0.028 |
| 25 | 5.721 | 5.729 | 0.008 | 6.058 | 0.056 |
| 26 | 5.638 | 5.595 | −0.043 | 6.049 | 0.068 |
| 27 | 6.319 | 6.331 | 0.012 | 6.837 | 0.076 |
| 28 | 6.509 | 6.538 | 0.029 | 6.164 | −0.056 |
| 29 | 6.167 | 6.186 | 0.019 | 6.167 | 0.000 |
| 30 | 6.065 | 6.148 | 0.083 | 6.195 | 0.021 |
| 31 | 6.468 | 6.485 | 0.017 | 6.082 | −0.063 |
| 32 | 6.959 | 7.152 | 0.193 | 6.968 | 0.001 |
| 33 | 7.301 | 7.338 | 0.037 | 7.300 | 0.000 |
| 34 | 8.155 | 8.073 | −0.082 | 7.161 | −0.139 |
| 35 | 7.260 | 7.270 | 0.010 | 7.108 | −0.021 |
| 36 | 6.921 | 6.812 | −0.109 | 6.929 | 0.001 |
| 37 | 6.959 | 6.998 | 0.039 | 7.103 | 0.020 |
| 38 | 7.745 | 7.733 | −0.012 | 7.744 | 0.000 |
| 39 | 7.174 | 7.123 | −0.051 | 6.850 | −0.047 |
| 40 | 6.721 | 6.929 | 0.208 | 7.148 | 0.060 |
| 41 | 5.886 | 5.898 | 0.012 | 6.415 | 0.082 |
| 42 | 6.366 | 6.376 | 0.010 | 6.378 | 0.002 |
| 43 | 6.444 | 6.603 | 0.159 | 6.461 | 0.003 |
| 44 | 5.854 | 5.796 | −0.058 | 6.354 | 0.079 |
| 45 | 5.886 | 5.880 | −0.006 | 6.106 | 0.036 |
| 46 | 5.347 | 5.519 | 0.172 | 5.649 | 0.053 |
| 47 | 6.824 | 6.387 | −0.437 | 6.829 | 0.001 |
| 48 | 6.456 | 6.432 | −0.024 | 7.148 | 0.097 |
| 49 | 7.276 | 7.092 | −0.184 | 7.148 | −0.018 |
| 50 | 6.076 | 6.145 | 0.069 | 6.674 | 0.090 |
| 51 | 6.523 | 6.631 | 0.108 | 7.075 | 0.078 |
| 52 | 7.244 | 7.403 | 0.159 | 7.227 | −0.002 |
| 53 | 6.886 | 6.903 | 0.017 | 6.798 | −0.013 |
| 54 | 6.468 | 6.494 | 0.026 | 6.738 | 0.040 |
| 55 | 7.076 | 7.071 | −0.005 | 6.939 | −0.020 |
| 56 | 7.091 | 7.209 | 0.118 | 7.090 | 0.000 |
| 57 | 6.733 | 6.729 | −0.004 | 7.185 | 0.063 |
| 58 | 6.971 | 6.945 | −0.026 | 6.692 | −0.042 |
| 59 | 7.678 | 7.509 | −0.169 | 6.897 | −0.113 |
| 60 | 6.288 | 6.289 | 0.001 | 5.990 | −0.050 |
| 61 | 6.197 | 6.251 | 0.054 | 6.727 | 0.079 |
| 62 | 6.824 | 6.786 | −0.038 | 6.988 | 0.023 |
| 63 | 7.638 | 7.712 | 0.074 | 7.589 | −0.006 |
| 64 | 7.523 | 7.528 | 0.005 | 7.523 | 0.000 |
| 65 | 8.155 | 8.012 | −0.143 | 8.058 | −0.012 |
| 66 | 7.523 | 7.403 | −0.120 | 7.523 | 0.000 |
| 67 | 6.468 | 6.463 | −0.005 | 6.497 | 0.004 |
| 68 | 7.046 | 7.065 | 0.019 | 7.229 | 0.025 |
| 69 | 7.337 | 7.351 | 0.014 | 7.337 | 0.000 |
| 70 | 8.187 | 8.157 | −0.030 | 8.139 | −0.006 |
| 71 | 7.959 | 7.982 | 0.023 | 7.217 | −0.103 |
| 72 | 8.000 | 8.013 | 0.013 | 7.938 | −0.008 |
| 73 | 8.155 | 8.125 | −0.030 | 8.156 | 0.000 |
| 74 | 8.155 | 8.158 | 0.003 | 8.156 | 0.000 |
| 75 | 7.620 | 7.641 | 0.021 | 7.638 | 0.002 |
| 76 | 8.155 | 8.309 | 0.154 | 8.155 | 0.000 |

Table 2 (*Continued*)

| Compound no. | $pIC_{50}$ observed | $pIC_{50}$ predicted (PLS) | REP | $pIC_{50}$ predicted (PC-RBFNN) | REP |
|---|---|---|---|---|---|
| 77 | 7.553 | 7.580 | 0.027 | 7.936 | 0.048 |
| 78 | 8.301 | 8.279 | −0.022 | 7.553 | −0.099 |
| 79 | 7.796 | 7.819 | 0.023 | 7.315 | −0.066 |
| 80 | 8.097 | 8.101 | 0.004 | 8.098 | 0.000 |
| 81 | 7.222 | 7.423 | 0.201 | 7.276 | 0.007 |
| 82 | 7.886 | 7.926 | 0.040 | 7.809 | −0.010 |
| 83 | 7.569 | 7.584 | 0.015 | 7.694 | 0.016 |
| 84 | 8.155 | 7.999 | −0.156 | 7.570 | −0.077 |
| 85 | 8.046 | 8.032 | −0.014 | 7.932 | −0.014 |
| 86 | 8.046 | 8.067 | 0.021 | 8.047 | 0.000 |
| 87 | 7.244 | 7.243 | −0.001 | 8.217 | 0.118 |
| 88 | 8.301 | 8.319 | 0.018 | 7.245 | −0.146 |
| 89 | 6.620 | 6.607 | −0.013 | 7.928 | 0.165 |
| 90 | 8.155 | 7.989 | −0.166 | 7.217 | −0.130 |
| 91 | 7.921 | 7.743 | −0.178 | 7.196 | −0.101 |
| 92 | 8.000 | 7.992 | −0.008 | 7.216 | −0.109 |
| 93 | 7.180 | 7.110 | −0.070 | 7.156 | −0.003 |
| 94 | 5.939 | 5.944 | 0.005 | 6.172 | 0.038 |

**Table 3**
Statistics parameters and figures of merits of developed models.

| Statistics | PLS | | PC-RBFNN | |
|---|---|---|---|---|
| | Training set | Test set | Training set | Test set |
| $N$ | 73 | 20 | 73 | 20 |
| $R^2$ | 0.991 | 0.958 | 0.936 | 0.946 |
| *RMSE* | 0.092 | 0.145 | 0.221 | 0.212 |
| *PRESS* | 0.473 | 0.418 | 3.601 | 0.899 |
| $R^2_{CV}$ | 0.993 | | 0.939 | |
| $RMSE_{CV}$ | 0.100 | | 0.254 | |
| $R^2 - R_0^2/R^2$ | −0.009 | −0.044 | −0.065 | −0.038 |
| $R^2 - R'^2_0/R^2$ | −0.009 | −0.044 | −0.064 | −0.040 |
| $k$ | 1.000 | 0.999 | 1.006 | 1.013 |
| $k'$ | 1.000 | 1.000 | 0.993 | 0.986 |
| $R^2_m$ | 0.900 | 0.761 | 0.705 | 0.765 |

is generally tested using test set and leave one out cross validation. The satisfactory prediction of values of inhibitory activity of test set compounds demonstrates the efficacy of the MIA-QSAR in predicting activities of external molecules. Moreover, the low values of *RMSE* and predicted error sum of square (*PRESS*) for the prediction of activity of molecules in test set increases the statistical significance of the developed model. Also it must be noticed that the developed model passed all Tropsha and Roy parameters indicating high degree of predictability of the developed model.

Applying *Y*-randomization, the developed PLS model was further validated. Several random shuffles of the *Y* vector were performed and the results are shown in Table 4. The low $R^2$ values show that the good results in our original model are not due to a chance correlation or structural dependence of the training set. Also *F* statistic was calculated and its value was 76.332.

**Table 4**
Squared correlation coefficients obtained in two models by *Y* randomization.

| Iteration | $R^2$ | |
|---|---|---|
| | PLS | PC-RBFNN |
| 1 | 0.021 | 0.201 |
| 2 | 0.119 | 0.292 |
| 3 | 0.213 | 0.122 |
| 4 | 0.016 | 0.239 |
| 5 | 0.119 | 0.061 |
| 6 | 0.225 | 0.291 |
| 7 | 0.241 | 0.193 |
| 8 | 0.032 | 0.131 |
| 9 | 0.203 | 0.330 |
| 10 | 0.079 | 0.291 |

### 3.2. PC-RBFNN

As it was discussed above, radial basis neural network was chosen for constructing non linear model. To avoid the problem of collinearity, PCs of original descriptors were used. If the selected number of $n_h$ is lower than optimum number, the derived model is called underfitted model and may not calculate true activity of molecules. On the other hand, if too many hidden layer units are used the network is overfitted. Thus for initial training of network, we chose 10 hidden nodes and the spread equal to 1.0 and these values were used for finding optimum number of PC-RBFNN components. This optimization was performed by jointly analyzing *RMSEC* and *RMSECV*. As it is shown in Fig. 3B, 3PCs were selected as optimum number of PCs. The performance of PC-RBFNN model is significantly influenced by parameters of networks namely the number of radial basis functions $n_h$ and spread of networks (Fig. 6).

With 3 PCs, a response surface methodology was used to optimize $n_h$ and spread of networks. As is shown in Fig. 5 the surface plot of *RMSECV* as a function of $n_h$ and spread was plotted. $n_h$ was changed from 1 to 70 and spread from 0.1 to 3 in increments of 0.1. These ranges were selected according to the previous studies. The results show that a PC-RBFNN with 27 nodes in hidden layer and spread of 0.7 resulted in the optimum network performance.

The developed model was trained using the data of training set and it was evaluated by test samples. The predicted values of inhibitory activity of the studied compounds resulted from the optimized PC-RBFNN procedures are reported in Table 2, in association with relative error of prediction and are plotted in Fig. 4B against their corresponding experimental values. The statistical parameters and figures of merit as well as Tropsha and Roy parameters for determining the predictability of the developed model are presented for the best-fitted model in Table 3. As presented in Table 3 the model gave an *RMSE* of 0.092 for the training set and 0.145 for the test set, and the corresponding correlation coefficient $R^2$ of 0. 936 and 0.946 respectively. The *LOO* cross-validated coefficient was 0.939 for training set. As can be seen in Table 3 comparison of the PLS and PC-RBFNNs shows that the PLS model can simulate the relationship between the structure descriptors obtained by multivariate imaging analysis and the activity of studied cyclin dependent kinase 4 inhibitors more accurately. Furthermore, on the basis of criteria recommended by Tropsha and also $R^2_m$ by Roy, the obtained model is very predictive.

The PC-RBFNN was further validated by applying the *Y*-randomization test. In particular, 10 random shuffles of the *Y*-vector gave low $R^2$ values. These values are reported in Table 4. This shows that the developed PC-RBFNN model was not obtained by chance.

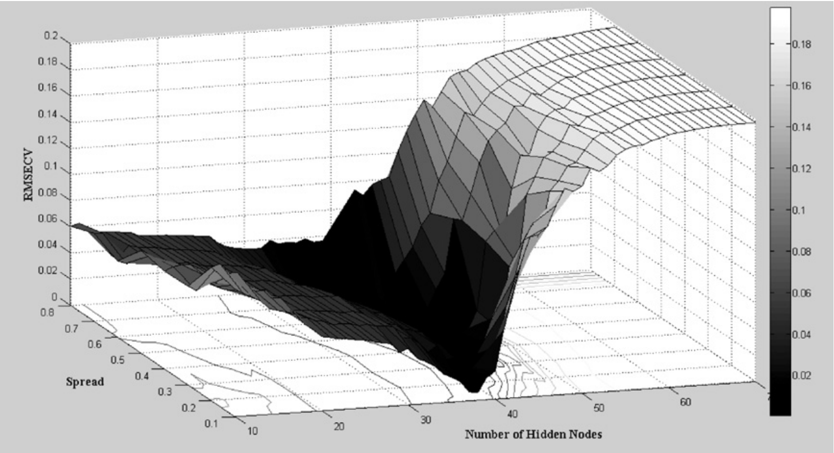


**Fig. 6.** Optimization of number of hidden nodes and spread value in PC-RBFNN model.

As a final point, one could dispute that how researchers can interpret the developed models by MIA-QSAR or how MIA-QSAR can be applied to propose novel compounds with improved activity. Said another way, what does the developed models mean to medicinal chemists? As discussed above, the calculated descriptors in MIA-QSAR do not mean physicochemically, 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 he can use the developed MIA-QSAR model to predict the activity of his proposed molecules. This practice may lead to the introduction of biologically active molecules.

Results of sensitivity (SE), specificity (SP), accuracy (ACC), and Matthew's correlation coefficient (MCC) for PC-RBFNN were also calculated. The PC-RBFNN model revealed a good fitted model with respect to these parameters (accuracy = 97.01%, sensitivity = 0.946, specificity = 0.934 and MCC = 0.911).

## 4. Conclusions

In this study, we have demonstrated the detailed application of multivariate image analysis method for the evaluation of quantitative structure activity relationship of some cyclin dependent kinase 4 inhibitors. Since in MIA-QSAR method descriptors are pixels of bitmaps of molecules, and because of high number of descriptors and potential problem of collinearity between them we used components produced by principal component analysis and partial least squares. Models were built after developing new orthogonal descriptors. The performance of developed models namely PLS and PC-RBFNNs were tested by several validation methods such as external and internal tests and also criteria recommended by Tropsha and Roy. The resulted PLS model had a high statistical quality ($R^2 = 0.991$ and $R^2_{CV} = 0.993$) for predicting the activity of the compounds. For PC-RBFNNs model these values were $R^2 = 0.936$ and $R^2_{CV} = 0.939$. Because of high correlation between values of predicted and experimental activities, MIA-QSAR proved to be a highly predictive approach. Comparison between predictability of PLS and PC-RBFNNs indicates that linear method has a higher power than nonlinear method in the prediction of activity of studied cyclin dependent kinase 4 inhibitors.

## References

[1] T. Xie, Y. Niu, K. Ge, S. Lu, Regulation of keratinocyte proliferation in rats with deep, partial-thickness scald: modulation of cyclin D1-cyclin-dependent kinase 4 and histone H1 kinase activity of M-phase promoting factor, J. Surg. Res. 147 (2008) 9–14.

[2] D. Fabbro, C. Garcia-Echeverria, Targeting protein kinases in cancer therapy, Curr. Opin. Drug. Discov. Dev. 5 (2002) 701–712.

[3] L. Zhu, Tumour suppressor retinoblastoma protein Rb: a transcriptional regulator, Eur. J. Cancer 41 (2005) 2415–2427.

[4] J. Bartek, J. Bartkova, J. Lukas, The retinoblastoma protein pathway and the restriction point, Curr. Opin. Cell Biol. 8 (1996) 805–814.

[5] A.S. Lundberg, R.A. Weinberg, Functional inactivation of the retinoblastoma protein requires sequential modification by at least two distinct cyclin–cdk complexes, Mol. Cell. Biol. 18 (1998) 753–761.

[6] R.A. Weinberg, The retinoblastoma protein and cell cycle control, Cell 81 (1995) 323–330.

[7] L. Yamasaki, Role of the RB tumor suppressor in cancer, Cancer. Treat. Res. 115 (2003) 209–239.

[8] S. Ortega, M. Malumbres, M. Barbacid, Cyclin D-dependent kinases, INK4 inhibitors and cancer, Biochem. Biophys. Acta 1602 (2002) 73–87.

[9] A.M. Senderowicz, E.A. Sausville, Preclinical and clinical development of cyclin-dependent kinase modulators, J. Natl. Cancer Inst. 92 (2000) 376–387.

[10] M. Knockaert, P. Grrengard, L. Meijer, Pharmacological inhibitors of cyclin-dependent kinases, Trends Pharmacol. Sci. 23 (2002) 417–425.

[11] E.A. Sausville, Cyclin-dependent kinase modulators studied at the NCI: pre-clinical and clinical studies, Curr. Med. Chem. Anti-cancer Agents 3 (2003) 47–56.

[12] P.M. Fischer, A. Gianella-Borradori, Recent progress in the discovery and development of cyclin-dependent kinase inhibitors, Expert Opin. Investig. Drugs 14 (2005) 457–477.

[13] M. Malumbres, M. Barbacid, Is cyclin D1-CDK4 kinase a bona fide cancer target? Cancer Cell 9 (2006) 2–4.

[14] L. Eriksson, S. Wold, J. Trygg, Multivariate analysis of congruent images, J. Chemometr. 19 (2005) 393–403.

[15] M. Shahlaei, R. Sabet, M.B. Ziari, B. Moeinifard, A. Fassihi, R. Karbakhsh, 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. 45 (2010) 4499–4508.

[16] M. Shahlaei, A. Fassihi, L. Saghaie, Application of PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: a comparative study, Eur. J. Med. Chem. 45 (2010) 1572–1582.

[17] M. Pompe, M. Razinger, M. Novič, M. Veber, Modelling of gas chromatographic retention indices using counterpropagation neural networks, Anal. Chim. Acta 348 (1997) 215–221.

[18] B. Walkzak, D.L. Massart, Local modelling with radial basis function networks, Chemom. Intell. Lab. Syst. 50 (2000) 179–198.

[19] J. Tetteh, S. Howells, E. Metcalfe, T. Suzuki, Optimisation of radial basis function neural networks using biharmonic spline interpolation, Chemom. Intell. Lab. Syst. 41 (1998) 17–29.

[20] D.A. Nugiel, A.-M. Etzkorn, A. Vidwans, P.A. Benfield, M. Boisclair, C.R. Burton, S. Cox, P.M. Czerniak, D. Doleniak, S.P. Seitz, Indenopyrazoles as novel cyclin dependent kinase (CDK) inhibitors, J. Med. Chem. 44 (2001) 1334–1336.

[21] D.A. Nugiel, A. Vidwans, A.-M. Etzkorn, K.A. Rossi, P.A. Benfield, C.R. Burton, S. Cox, D. Doleniak, S.P. Seitz, Synthesis and evaluation of indenopyrazoles as cyclin-dependent kinase inhibitors. 2. Probing the indeno ring substituent pattern, J. Med. Chem. 45 (2002) 5224–5232.

[22] E.W. Yue, C.A. Higley, S.V. DiMeo, D.J. Carini, D.A. Nugiel, C. Benware, P.A. Benfield, C.R. Burton, S. Cox, R.H. Grafstrom, D.M. Sharp, L.M. Sisk, J.F. Boylan, J.K. Muckelbauer, A.M. Smallwood, H. Chen, C.-H. Chang, S.P. Seitz, G.L. Trainor, Synthesis and evaluation of indenopyrazoles as cyclin-dependent kinase inhibitors. 3. Structure activity relationships at C3[1,2], J. Med. Chem. 45 (2002) 5233–5248.

[23] E.W. Yue, S.V. DiMeo, C.A. Higley, J.A. Markwalder, C.R. Burton, P.A. Benfield, R.H. Grafstrom, S. Cox, J.K. Muckelbauer, A.M. Smallwood, H. Chen, C.-H. Chang,

G.L. Trainor, S.P. Seitz, Synthesis and evaluation of indenopyrazoles as cyclin-dependent kinase inhibitors. Part 4: Heterocycles at C3, Bioorg. Med. Chem. Lett. 14 (2004) 343–346.

[24] D.A. Nugiel, A. Vidwans, C.D. Dzierba, Parallel synthesis of acylsemicarbazide libraries: preparation of potent cyclin dependent kinase (cdk) inhibitors, Bioorg. Med. Chem. Lett. 14 (2004) (2004) 5489–5491.

[25] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute. Essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.

[26] R.W. Kennard, L.A. Stone, Computer-aided design of experiments, Technometrics 11 (1969) 137–149.

[27] Y. Fan, L.M. Shi, K.W. Kohn, Y. Pommier, J.N. Weinstein, Quantitative structure–antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies, J. Med. Chem. 44 (2001) 3254–3326.

[28] X.J. Yao, A. Panaye, P. Doucet, R.S. Zhang, H.F. Chen, M.C. Liu, Z.D. Hu, B.T. Fan, Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression, J. Chem. Inf. Comput. Sci. 44 (2004) 1257–1266.

[29] J. Shi, F. Luan, H.X. Zhang, M.C. Liu, Q.X. Guo, Z.D. Hu, B.T. Fan, QSPR study of fluorescence wavelengths (lambda ex/lambda em) based on the heuristic method and radial basis function neural networks, QSAR Comb. Sci. 25 (2006) 147–155.

[30] A. Golbraikh, A. Tropsha, Beware of $q^2$, J. Mol. Grap. Model. 20 (2002) 269–276.

[31] P. Roy, K. Roy, On some aspects of variable selection for partial least squares. Regression models, QSAR Comb. Sci. 27 (2008) 302–313.