

## QSAR study of some 5-methyl/trifluoromethoxy-1H-indole-2,3-dione-3-thiosemicarbazone derivatives as anti-tubercular agents

M. Shahlaei<sup>1,2</sup>, A. Fassihi<sup>2,3,\*</sup> and A. Nezami<sup>4</sup>

<sup>1</sup>Department of Medicinal Chemistry, Faculty of Pharmacy, Kermanshah University of Medical Sciences, Kermanshah, I.R.Iran.

<sup>2</sup>Department of Medicinal Chemistry, School of Pharmacy and Pharmaceutical Sciences, Isfahan University of Medical Sciences, Isfahan, I.R.Iran.

<sup>3</sup>Isfahan Pharmaceutical Sciences Research Center, School of Pharmacy and Pharmaceutical Sciences, Isfahan University of Medical Sciences, Isfahan, I.R.Iran.

<sup>4</sup>Department of Pediatrics, Faculty of Medicine, Lorestan University of Medical Sciences, Khoramabad, I.R.Iran.

---

### Abstract

In the present study, quantitative relationships between molecular structure and anti-tubercular activity of some 5-methyl/trifluoromethoxy-1H-indole-2,3-dione-3-thiosemicarbazone derivatives were discovered. The detailed application of an efficient linear method and principal component regression (PCR) for the evaluation of quantitative structure activity relationships of the studied compounds is demonstrated. Components produced by principal component analysis were used as the input for a linear model development. Results indicate a linear relationship between the principal components obtained from molecular descriptors and the inhibitory activity of this set of molecules. The maximum variance in the activity of the molecules in PCR method was 73%. The performance of the developed model was tested by several validation methods.

**Keywords:** QSAR; Thiosemicarbazone derivaives; Principal component analysis; Anti-tuberculosis activity

---

### INTRODUCTION

Tuberculosis (TB) is an infectious condition with high degree of mortality all over the world. According to the World Health Organization one-third of the world's population was infected with *Mycobacterium tuberculosis* in 2006 (1). Other statistics indicate that 8 million of new cases are estimated to appear annually, 95% of which occurs in developing countries, and almost 2 million of the sufferers are believed to die from the disease (2,3). A large number of the infected people bear the latent form of TB which produces a possible risky source of the sickness for the future generations. The HIV epidemic has resulted in the fast increase of the TB pandemic and has enhanced the probability of death from TB. The emergence of multidrug resistant tuberculosis to the first-

line drugs such as isoniazid, rifampicin, ethambutol, streptomycin and pyrazinamide has made the disease hard to cure (4-6). This obstacle in the treatment of TB and the statistical facts about its prevalence highlights the necessity of designing novel, more potent and less prone to resistance compounds with fewer side effects.

The design of new anti-tubercular drugs may be accomplished using the sophisticated computational methods which are generally based on two different approaches. Designing new ligands for inhibiting a known biochemical target by noticing its structural features which seems a more ideal approach and proposing novel biologically active compounds after statistical analyses of the structures of some known drugs having the desired biological activities. The first approach is called structure-based drug design, and the

---

\*Corresponding author: Afshin Fassihi, Pharm.D; Ph.D  
Tel. 0098 311 7922562, Fax. 0098 311 6680011  
Email: fassihi @pharm.mui.ac.ir

other is referred to as ligand-based drug design (LBDD) (7). Quantitative structure activity relationship (QSAR) is a method of the LBDD approach. As one of the most powerful techniques for predicting the bioactivity of various molecules, QSAR starts only with the molecular structure information of the previously reported active molecules. Such a data mining approach offers important insight into the relationships between the molecular structure and biological activity of the investigated compounds by means of statistical models. Various methods have been applied to construct QSAR models including linear and nonlinear regression methods (8,9).

Here we investigate the quantitative structure activity relationships of a series of 49 methyl/trifluoromethoxy 1*H*-indole-2,3-dione-3-thiosemicarbazone derivatives reported in

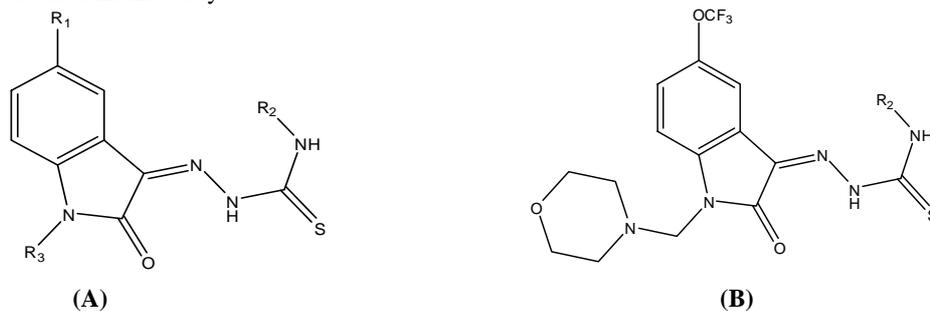
literature as anti-tubercular compounds. Principal component regression (PCR) as a method for feature extraction was employed, and its output was used as the input of linear regression. After reaching the model it was validated using various model validation techniques.

## MATERIALS AND METHODS

### Data preparation

All calculations were performed in a Pentium IV personal computer (CPU at 2.6 GHz) with Windows XP operating system. All biological data used here were derived from the literature (10). General structures and the structural details of these compounds are reported in Table 1.

**Table 1.** General structures and structural details of 5-methyl/trifluoromethoxy-1*H*-indole-2,3-dione-3-thiosemicarbazone derivatives used in this study.



Compound	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	Type
1	CH <sub>3</sub>	CH <sub>3</sub> CH=CH <sub>2</sub>	H	A
2	CH <sub>3</sub>	C <sub>4</sub> H <sub>9</sub>	H	A
3	CH <sub>3</sub>	C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub>	H	A
4	CH <sub>3</sub>	4-FC <sub>6</sub> H <sub>4</sub>	H	A
5 <sup>a</sup>	CH <sub>3</sub>	2-BrC <sub>6</sub> H <sub>4</sub>	H	A
6	CH <sub>3</sub>	3-BrC <sub>6</sub> H <sub>4</sub>	H	A
7	CF <sub>3</sub> O	4-NO <sub>2</sub> C <sub>6</sub> H <sub>4</sub>	H	A
8 <sup>a</sup>	CF <sub>3</sub> O	CH <sub>3</sub>	H	A
9	CF <sub>3</sub> O	C <sub>2</sub> H <sub>5</sub>	H	A
10	CF <sub>3</sub> O	CH <sub>3</sub> CH=CH <sub>2</sub>	H	A
11	CF <sub>3</sub> O	C <sub>4</sub> H <sub>9</sub>	H	A
12	CF <sub>3</sub> O	<i>cycl</i> -C <sub>6</sub> H <sub>11</sub>	H	A
13 <sup>a</sup>	CF <sub>3</sub> O	C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub>	H	A
14	CF <sub>3</sub> O	C <sub>6</sub> H <sub>5</sub>	H	A
15 <sup>a</sup>	CF <sub>3</sub> O	4-CH <sub>3</sub> C <sub>6</sub> H <sub>4</sub>	H	A
16	CF <sub>3</sub> O	4-CH <sub>3</sub> OC <sub>6</sub> H <sub>4</sub>	H	A
17	CF <sub>3</sub> O	4-FC <sub>6</sub> H <sub>4</sub>	H	A
18	CF <sub>3</sub> O	4-ClC <sub>6</sub> H <sub>4</sub>	H	A
19	CF <sub>3</sub> O	4-BrC <sub>6</sub> H <sub>4</sub>	H	A

Table 1. (continued)

Compound	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	Type
20	CH <sub>3</sub>	2-BrC <sub>6</sub> H <sub>4</sub>	CH <sub>3</sub>	A
21	CF <sub>3</sub> O	C <sub>2</sub> H <sub>5</sub>	CH <sub>3</sub>	A
22	CF <sub>3</sub> O	CH <sub>3</sub> CH=CH <sub>2</sub>	CH <sub>3</sub>	A
23	CF <sub>3</sub> O	C <sub>4</sub> H <sub>9</sub>	CH <sub>3</sub>	A
24	CF <sub>3</sub> O	<i>cycl</i> -C <sub>6</sub> H <sub>11</sub>	CH <sub>3</sub>	A
25	CF <sub>3</sub> O	C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub>	CH <sub>3</sub>	A
26 <sup>a</sup>	CF <sub>3</sub> O	4-CH <sub>3</sub> C <sub>6</sub> H <sub>4</sub>	CH <sub>3</sub>	A
27	CF <sub>3</sub> O	4-FC <sub>6</sub> H <sub>4</sub>	CH <sub>3</sub>	A
28	CF <sub>3</sub> O	4-ClC <sub>6</sub> H <sub>4</sub>	CH <sub>3</sub>	A
29	CF <sub>3</sub> O	4-BrC <sub>6</sub> H <sub>4</sub>	CH <sub>3</sub>	A
30	-	CH <sub>3</sub>	-	B
31 <sup>a</sup>	-	C <sub>2</sub> H <sub>5</sub>	-	B
32	-	CH <sub>3</sub> CH=CH <sub>2</sub>	-	B
33	-	C <sub>4</sub> H <sub>9</sub>	-	B
34	-	<i>cycl</i> -C <sub>6</sub> H <sub>11</sub>	-	B
35 <sup>a</sup>	-	C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub>	-	B
36	-	C <sub>6</sub> H <sub>5</sub>	-	B
37 <sup>a</sup>	-	4-CH <sub>3</sub> C <sub>6</sub> H <sub>4</sub>	-	B
38	-	4-CH <sub>3</sub> OC <sub>6</sub> H <sub>4</sub>	-	B
39	-	4-FC <sub>6</sub> H <sub>4</sub>	-	B
40	-	4-ClC <sub>6</sub> H <sub>4</sub>	-	B
41	-	4-BrC <sub>6</sub> H <sub>4</sub>	-	B
42	-	4-NO <sub>2</sub> C <sub>6</sub> H <sub>4</sub>	-	B
43	CH <sub>3</sub>	C <sub>2</sub> H <sub>5</sub>	H	A
44	CH <sub>3</sub>	<i>cycl</i> -C <sub>6</sub> H <sub>11</sub>	H	A
45 <sup>a</sup>	CH <sub>3</sub>	C <sub>6</sub> H <sub>5</sub>	H	A
46	CH <sub>3</sub>	4-CH <sub>3</sub> C <sub>6</sub> H <sub>4</sub>	H	A
47	CH <sub>3</sub>	4-ClC <sub>6</sub> H <sub>4</sub>	H	A
48	CH <sub>3</sub>	4-BrC <sub>6</sub> H <sub>4</sub>	H	A
49	CH <sub>3</sub>	C <sub>6</sub> H <sub>5</sub>	CH <sub>3</sub>	A

<sup>a</sup>Molecules assigned by Kennard and Stone algorithm as test set

pIC<sub>50</sub> (log 1/IC<sub>50</sub>) is the dependent variable that characterizes the biological parameter for the developed QSAR model. The structures of molecules were drawn and optimized using Hyperchem 7.0 software (11). Semi-empirical AM1 method with Polak-Ribiere algorithm until the root mean square gradient of 0.01 kcal/mol was the optimization method. The resulted geometries were transferred into Dragon program, version 2.1 (developed by Milano Chemometrics and QSAR Group) to calculate some descriptors (12). A large number of theoretical different descriptors were calculated for each molecule. The name and number of calculated descriptors can be seen in the Table 2. These four classes of

descriptors were selected to ease the calculations. All the required model development calculations were performed within the MATLAB (version 7.1, MathWorks, Inc.) environment.

In this study, the Kennard and Stone algorithm was employed for assigning the training and test sets (13). This method of data splitting has some advantages: the training set molecules map the measured region of the input variable space completely with respect to the induced metric. The other advantage is that the test molecules all fall inside the measured region.

Principal component analysis (PCA) was employed to compress a pool of calculated

**Table 2.** Some descriptors used in model building.

Descriptor	Molecular Descriptor
<b>Constitutional</b>	Molecular weight, no. of atoms, no. of non-H atoms, no. of bonds, no. of heteroatoms, no. of multiple bonds (nBM), no. of aromatic bonds, no. of functional groups (hydroxyl, amine, aldehyde, carbonyl, nitro, nitroso, etc.), no. of rings, no. of circuits, no of H-bond donors, no of H-bond acceptors, no. of Nitrogen atoms (nN), chemical composition, sum of Kier-Hall electrotopological states (Ss), mean atomic polarizability (Mp), number of rotatable bonds (RBN), etc.
<b>Topological</b>	Molecular size index, molecular connectivity indices (X1A, X4A, X2v, X1Av, X2Av, X3Av, X4Av), information content index (IC), Kier Shape indices, total walk count, path/walk-Randic shape indices (PW3, PW4, Zagreb indices, Schultz indices, Balaban J index (such as MSD) Wiener indices, topological charge indices, Sum of topological distances between F.F (T(F..F)), Ratio of multiple path count to path counts (PCR), Mean information content vertex degree magnitude (IVDM), Eigenvalue sum of Z weighted distance matrix (SEigZ), reciprocal hyperdetour index (Rww), Eigenvalue coefficient sum from adjacency matrix (VEA1), radial centric information index, 2D petijean shape index (PJI2), etc.
<b>Geometrical</b>	D petijean shape index (PJI3), Gravitational index, Balaban index, Wiener index, etc
<b>Functional group</b>	Number of total tertiary carbons (nCt), Number of H-bond acceptor atoms (nHAcc), number of total hydroxyl groups (nOH), number of unsubstituted aromatic C(nCaH), number of ethers (aromatic) (nRORPh), etc.

descriptors into principal components (PCs) as new variables.

As a matter of fact, PCR is a multiple linear regression method which uses the score matrix as new variables for building of the model.

After the building of the model, its validation is a crucial part of any QSAR procedure. In other words, after the calculation of the regression coefficients (b) by the least square methods, the coefficients are used to predict the activity of external test set. For example, if  $X_c$  and  $X_t$  are the matrices of factors for calibration and test sets respectively, the  $Y_c$  and  $Y_t$  matrices of activity for calibration and test sets can be obtained using the following equations:

$$Y_c = b X_c$$

$$Y_t = b X_t$$

Matrix of b can be calculated.

The statistical qualities of the generated QSAR models were evaluated using methods such as leave one out cross validation and parameters like predicted residual sum of squares (PRESS) and the root mean square error (RMSE).

Leave one out cross validation is a well-known and accepted method applied to discover the reliability of the generated QSAR models. Based on this method, a number of modified data sets (equals to the number of the studied molecules) are generated by removing one of the molecules in each case. For each

new data set, a model is generated using the modeling procedure applied in the study. Each model is examined through the evaluation of its power in predicting the bioactivity of deleted molecule. This process is repeated until a total set of predicted bioactivity for all of the investigated molecules is achieved. The predictive ability will be evaluated by the cross validation coefficient ( $R_{cv}^2$ ) calculated using the following equation:

$$R_{cv}^2 = 1 - \frac{\sum_{i=1}^n (y_{exp,i} - y_{pred,i})^2}{\sum_{i=1}^n (y_{exp,i} - \bar{y})^2}$$

Root mean square error of cross validation (RMSE<sub>CV</sub>) for the developed models is reported, as well

Some criteria for the prediction of the model are suggested by Tropsha. If these criteria are satisfied, it can then be concluded that the model is predictive (14).

These criteria include:

$$R_{LOO}^2 > 0.5 \quad R^2 > 0.6$$

$$\frac{R^2 - R_o^2}{R^2} < 0.1 \quad \frac{R^2 - R_o'^2}{R^2} < 0.1$$

$$0.85 < k < 1.15 \quad \text{or} \quad 0.85 < k' < 1.15$$

where,  $R^2$  is the correlation coefficient of regression between the predicted and observed activities of compounds in training and test set.

$R_o^2$  is the correlation coefficients for regressions between predicted versus observed activities through the origin,  $R_o'^2$  is the

**Table 3.** Eigenvalues of calculated PCs, % of explained variances and cumulative variances.

PC No.	Eigenvalue	% variance explained	cumulative variance
1	195.0	60.70	60.70
2	39.46	12.26	72.96
3	27.92	8.672	81.63
4	12.28	3.814	85.45
5	8.755	2.719	88.16
6	8.281	2.572	90.73
7	5.879	1.826	92.56
8	4.508	1.400	93.96
9	3.887	1.207	95.17

correlation coefficient for the regressions between observed versus predicted activities through the origin, and the slopes of the regression lines through the origin are assigned by  $k$  and  $k'$ , respectively. Details of the definitions of parameters such as  $R_o^2$ ,  $R_o'^2$ ,  $k$  and  $k'$  are presented in the literature (14).

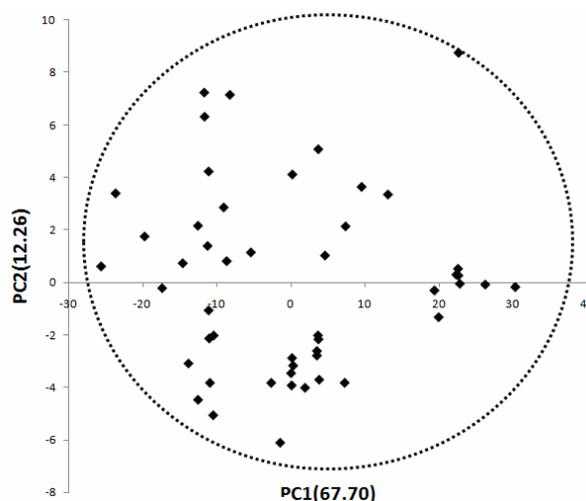
In addition, according to Roy and Roy (15), it is necessary to study the differences between the values of  $R_o^2$  and  $R_o'^2$ . They suggested the following modified  $R^2$  form:

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

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

## RESULTS

After deleting zero variance columns of X block, PCA was carried out on the pool of all descriptors. As evident in Table 3, among the generated PCs, only 9 eigenvalue ranked PCs were selected for the next model building. Table 3 demonstrates that PCA gives 9 significant PCs (% variance explained  $>1$ ) which can explain more than 95.17% of the variances in the original descriptors data matrix. Nine PCs with their eigenvalues are shown in the Table 3. In this Table, the eigenvalues, the percent of variances explained by each eigenvalue, and the cumulative percent of variances are represented. Therefore, the subsequent studies were restricted to these 9 PCs, and the selection of their best subset to perform the linear regression method. Plotting the first PC vs. the second one showed that none of



**Fig. 1** The first two components (PC1, and PC2) from the principal component analysis of the 49 considered molecules.

the compounds used in this study were outlier, although 5 clusters existed within the data set (Fig. 1).

When factor scores were used as the predictor parameters in a multiple regression equation using stepwise selection method of PCR, the following equation was obtained:

$$\begin{aligned} \text{pIC}_{50} &= 14.021 (\pm 2.143) + 12.210 (\pm 3.216) \\ &f1 + 1.453 (\pm 0.320) f3 \\ R^2 &= 0.83, \text{ S.E.} = 0.20, F = 20.05 \end{aligned}$$

where, F is F of ANOVA and S.E. is standard error in the resulted model.

The above equation shows high equation statistics (81% explained variance in  $\text{pIC}_{50}$  data). Since factor scores were used instead of selected descriptors, and each factor-score contains information from different descriptors, loss of information is avoided, and the

quality of PCR equation is better than similar equations such as those derived from MLR.

The cross-validation method was used to evaluate the robustness of the proposed model. In leave one out cross validation method, one object at a time was eliminated, and then PCR was performed on the remaining of training set. The activity of the left-out object was predicted using this regression model. This procedure was repeated until each compound in the calibration set was left out once. The optimum number of factors was selected with respect to the quantities of RMSECV, the root

mean square error of calibration, and 2 PCs were selected as the optimum number of PCs.

For the evaluation of the predictive power of the generated PCR, the optimized model was applied for the prediction of pIC<sub>50</sub> values of all compounds in the calibration and prediction set. The calculated pIC<sub>50</sub> for each molecule and relative error of prediction by model are summarized in Table 4. Very small values of relative errors confirm the accuracy of the proposed PCR model for modeling the anti-tubercular activity of the studied compounds.

**Table 4.** Calculated activities by PCR and their relative error of prediction (REP).

Compound	Experimental activity	Predicted activity	REP
1	3.743	3.853	0.029
2	5.379	5.124	-0.050
3	3.827	3.554	-0.077
4	4.866	4.983	0.024
5	4.538	4.667	0.028
6	4.941	4.776	-0.035
7	4.882	4.669	-0.046
8	5.115	5.555	0.079
9	5.197	5.544	0.063
10	4.262	4.554	0.064
11	3.478	3.854	0.097
12	4.955	4.459	-0.111
13	3.847	3.454	-0.114
14	4.459	4.836	0.078
15	4.476	4.433	-0.010
16	4.875	4.433	-0.100
17	4.900	4.654	-0.053
18	4.873	4.739	-0.028
19	4.424	4.763	0.071
20	4.260	4.654	0.085
21	4.223	4.530	0.068
22	4.564	4.674	0.024
23	3.390	3.557	0.047
24	3.704	3.640	-0.018
25	4.137	4.718	0.123
26	3.178	3.433	0.074
27	5.063	5.153	0.018
28	4.768	4.454	-0.071
29	5.213	5.460	0.045
30	4.805	4.406	-0.091
31	4.679	4.455	-0.050
32	4.999	4.652	-0.074
33	3.909	3.763	-0.039
34	4.741	4.630	-0.024

Table 4. (continued)

Compound	Experimental activity.	Predicted activity.	REP.
35	5.000	4.775	-0.047
36	4.315	4.634	0.069
37	4.862	4.394	-0.107
38	4.619	4.451	-0.038
39	4.633	4.123	-0.124
40	4.608	4.620	0.003
41	4.593	4.535	-0.013
42	4.175	4.285	0.026
43	4.157	4.436	0.063
44	5.152	5.439	0.053
45	4.968	4.529	-0.097
46	4.847	4.453	-0.089
47	4.598	4.440	-0.036
48	3.645	3.439	-0.060
49	3.332	3.237	-0.029

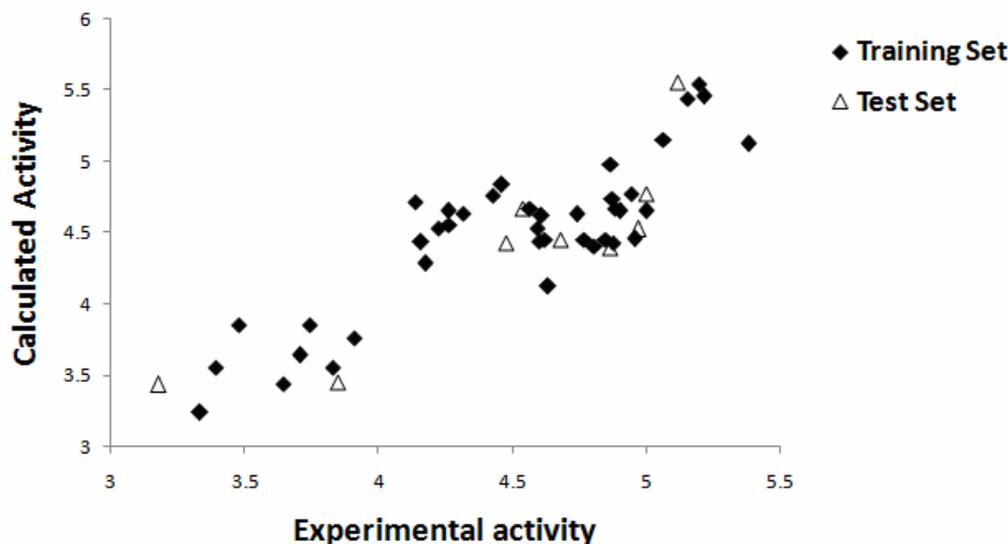


Fig. 2 Plot of calculated vs. experimental activity of investigated compounds in training and test sets.

Experimental versus predicted values for  $pIC_{50}$  values of training and test sets, obtained by the PCR modeling, are shown graphically in Fig. 2.

Results of various statistical criteria and figures of merit for this model for two subsets of molecules *i.e.* training and test sets are reported in Table 5. The external predictability of a proposed model is generally tested using test sets and  $R^2_{CV}$ . The satisfactory prediction of the values of the inhibitory activity of test set compounds demonstrates the efficacy of the QSAR in predicting the activities of

external molecules. Moreover, the low values of RMSE and PRESS for training, and test sets also add to the statistical significance of the developed models. Besides, on the basis of criteria recommended by Tropsha and also  $R_m^2$  by Roy, the obtained model is very predictive (Table 5).

**Table 5.** Statistical parameters obtained for the developed model for anti-tubercular activity of the investigated compounds.

Parameter	Training set	Test set
N	40	9
R <sup>2</sup>	0.735	0.762
RMSE	0.284	0.324
PRESS	3.274	0.946
R <sup>2</sup> <sub>LOOCV</sub>	0.792	
RMSE <sub>LOOCV</sub>	0.207	
R <sup>2</sup> <sub>L50.CV</sub>	0.746	
RMSE <sub>L50CV</sub>	0.255	
R <sup>2</sup> -R <sub>0</sub> <sup>2</sup> /R <sup>2</sup>	-0.008	-0.041
R <sup>2</sup> -R' <sub>0</sub> <sup>2</sup> /R <sup>2</sup>	-0.009	-0.041
k	1.023	0.987
k'	1.012	0.99
R <sub>m</sub> <sup>2</sup>	0.800	0.711
R <sup>2</sup> <sub>adjusted</sub>	0.713	

N: Number of objects in data set, R<sup>2</sup>: Correlation coefficient of experimental and predicted activities, RMSE: Root mean square error:  $1/N(\sqrt{\sum (y_i - \hat{y}_i)^2})$ , PRESS: Predicted error sum of square:  $\sqrt{\sum (y_i - \hat{y}_i)^2}$ , R<sup>2</sup><sub>cv</sub>: Correlation coefficient of leave one out cross validation, RMSE<sub>cv</sub>: Root mean square error of cross validation

## DISCUSSION

To obtain the effects of the structural features of the studied derivatives on their anti-tubercular activity, QSAR model development was performed with various calculated molecular descriptors (8). Because of the large number of calculated descriptors, PCA was employed to solve the collinearity problem in the generated descriptors, and the PCs were used as new variables for model building. In the PC analysis at first a data pre-processing step must be performed on the descriptors calculated by dragon using autoscaling. Suppose  $\mathbf{X}_{i,j}$  is the column mean-centered and scaled matrix of descriptors for  $i$  samples and  $j$  descriptors, and  $y_{i,1}$  the matrix of the activity (pIC<sub>50</sub>). After the generation of principal components, using matrix  $\mathbf{X}_{i,j}$ , the new matrix containing scores of PCs is created. Then these scores are used as new variables for regression. Scores as new variables possess two interesting properties (8): (i) They are sorted as the information content (variance) explaining decreases from the first PC to the last one of the PCs. As a result, the last PCs can be deleted, since they

don't have useful information. (ii) PCs are orthogonal, so the correlation problem that exists in the pool of descriptors calculated in this study is solved.

After the calculation of PCs, these factors are used as new variables in the building of the model. In order to evaluate the final developed model, the existing data set was divided into training and external prediction (test) sets. Almost 20% of the molecules (9 out of 49) were selected as external test set molecules. The training set plays an important role in developing the properties of the model.

The best situation in this stage of the model building is splitting of the data set to guarantee that both training set and test set individually cover the total space occupied by the original data set. The possibility of overfitting of the developed model is increased by the selection of more similar molecules as training set. Hence, ideal splitting of data set can be performed in such a way that each object in the test set is close to at least one of the objects in the training set. One of the best methods for data splitting is using Kennard and Stone algorithm. After dividing the molecules into two parts, training and test sets, based on

Kennard and Stone algorithm, building of the regression models using the calibration set was performed.

In order to get the linear relationship with independent variables, logarithms of the inverse of the biological activity ( $\log 1/IC_{50}$ ) data of 49 molecules were used.

An exact consideration of different statistical parameters indicated that the developed QSAR model could explain and predict 73% and 76% of the variances in the  $pIC_{50}$  in training and test sets data, respectively. It was observed that, the plot of data resulted by PCR represents the lowest scattering, with no systematic error. As shown in Table 5,  $R^2$  that is an indicative of the goodness of the fitting of the proposed model, was obtained for three sets, and the high value of this parameter indicates a good fitting between the PCs and the predicted values of anti-tubercular activities of the investigated compounds by developed PCR model. This shows the high predictability of the proposed model.

## CONCLUSION

Quantitative relationships between the molecular structure and the inhibitory activity of the series of some 5-methyl/trifluoromethoxy-1H-indole-2,3-dione-3-thiosemicarbazone derivatives as anti-tubercular agents were discovered by the collection of the calculated descriptors including topological, geometrical, constitutional, and functional group. As a result, it was found that correctly opted and designed PCR model could practically represent dependence of the 5-methyl/trifluoromethoxy-1H-indole-2,3-dione-3-thiosemicarbazone derivatives as anti-tubercular compounds to the extracted PCs from various geometrical, topological, and other calculated descriptors. The optimized principal regression method could simulate the linear relationship between  $pIC_{50}$  value and the PCs.

## REFERENCES

1. World Health Organization. Tuberculosis Fact sheet. Geneva: World Health Organization; 2006.
2. Frieden TR, Sterling TR, Munsiff SS, Watt CJ, Dye C. Tuberculosis. *Lancet*. 2003;362:887-899.
3. Snider DE, La Montagne JR. The neglected global tuberculosis problem: a report of the 1992 World Congress on Tuberculosis. *J Infect Dis*. 1994; 169:1189.
4. Sriram D, Yogeewari P, Madhu K. Synthesis and *in vitro* and *in vivo* antimycobacterial activity of Isonicotinoyl Hydrazones. *Bioorg Med Chem Lett*. 2005;15:4502-4505.
5. Zhang Y, Post-Martens K, Denkin S. New drug candidates and therapeutic targets for tuberculosis therapy. *Drug Discov Today*. 2006;11:21-27.
6. Duncan K, Barry CE. Prospects for new anti-tubercular drugs. *Curr Opin Microbiol*. 2004;7:460-465.
7. Young DC. Computational drug design. New Jersey: John Wiley; 2009.
8. 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.
9. Arkan E, Shahlaei M, Pourhossein A, Fakhri K, Fassihi A. Validated QSAR analysis of some diaryl substituted pyrazoles as CCR2 inhibitors by various linear and nonlinear multivariate chemometrics methods. *Eur J Med Chem*. 2010;45:3394-3406.
10. Güzel Ö, Karalı N, Salman A. Synthesis and antituberculosis activity of 5-methyl/trifluoromethoxy-1H-indole-2,3-dione-3-thiosemicarbazone derivatives, *Bioorg Med Chem*. 2008;16:8976-8987.
11. HyperChem Release 7.0 for Windows, Molecular modeling system: HyperCube. 2002. Available from: URL:<http://www.hyperchem.com>
12. Todeschini R. Milano chemometrics and QSPR group. 2010. Available from URL:<http://michem.disat.unimib.it/>
13. Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics*. 1969;11:137-149.
14. Tropsha A, Gramatica P, Gombar VK. 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.
15. Roy P, Roy K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb Sci*. 2008;27:302-313.