World Journal of Pharmaceutical and Life Sciences
WJPLS
www.wjpls.org

PREDICTIVE COMPARATIVE QSAR ANALYSIS OF AS NITROTRIAZOLE- AND IMIDAZOLE-BASED AMIDES DERIVATIVES MYCOBACTERIUM TUBERCULOSIS H37RV INHIBITORS

Mahesh. B. Palkar[1] and Chanabasayya. M. Vastrad[2*]

[1]Department of Pharmaceutical Chemistry K.L.E University, College of Pharmacy Vidyanagar, HUBLI – 580031, Karnataka, India.

[2]Department of Computer Science Mangalore University, Mangalagangotri-574 199, Karnataka, India.

Article Received on 10/05/2015 **Article Revised on 03/06/2015** **Article Accepted on 26/06/2015**

***Correspondence for Author**
Chanabasayya .M. Vastrad
Department of Computer Science Mangalore University,
Mangalagangotri- 574 199,
Karnataka, India.
channu.vastrad@gmail.com
palkarmahesh4u@rediffmail.com

ABSTRACT

Antitubercular activities of Nitrotriazole- and imidazole-based amides Derivatives series were subjected to Quantitative Structure Activity Relationship (QSAR) Analysis with an effort to derive and understand a correlation between the biological activity as response variable and different molecular descriptors as independent variables. QSAR models are built using 40 molecular descriptors dataset. Different statistical regression expressions were got using Partial Least Squares (PLS), Multiple Linear Regression (MLR) and Principal Component Regression (PCR) techniques. The among these technique, Multiple Linear Regression (MLR) technique has shown very promising result as compared to PLS technique A QSAR model was build by a training set of 30 molecular structures with correlation coefficient (r^2) *of 0.8340 , significant cross validated correlation coefficient* (q^2) *is 0.8123,* $F\ test$ *is 28.1025,* r^2 *for external test set* ($pred\_r^2$) *is 0.7945, coefficient of correlation of predicted data set* ($pred\_r^2se$) *is 0.5678 and degree of freedom is 22 by Multiple Linear Regression Technique.*

KEYWORDS: *TB, MLR, PLS, PCR, LOO.*

1. INTRODUCTION

Tuberculosis in humans is generally caused by mycobacterium tuberculosis(TB). The desease is spread by respirable droplets generated during effective expiratory manoeuvres such as coughing. TB desease can be either active or latent.[1] The World Health Organization (WHO) asses that within the next twenty years about thirty million people will be troubled with the bacillus.[2-3] The analytic management of TB has depends dully on a limited number of drugs such as Isonicotinic acid, Hydrazide, Rifadin, Rimactane, Myambutol, Streptomycin, Ethionamide, Pyrazinamide, Fluroquinolones etc.[4] Still with the origin of these special chemical drugs the spread of TB has not been eradicated completely because of delayed treatment programmes .There is now recognition that new drugs to treat TB are necessarily required, particularly for use in shorter medication procedure than are possible with the current agents and which can be engaged to treat multi-drug resistant and hidden disease.[5]

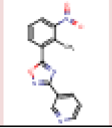
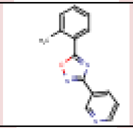
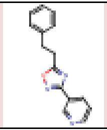
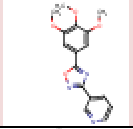
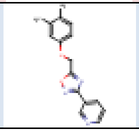
Nitrotriazole- and imidazole-based amides shows effective in vitro and in vivo antimycobacterial activity.[6] There is also a great effort to find and develop newer, 5-nitrofuran-2-yl, and some of them might have value in the remedy of TB.[7] Chemoinformatics[26] and computer-aided drug design (CADD) are likely to contribute to a possible solution for the dangerous situation regarding this infectious disease by assisting in the swift identification of new effective anti-TB agents. The other way for overcoming the absence of empirical analysis for biological systems is depends on the activity to develop quantitative structure activity relationship (QSAR).[8] QSAR models are mathematical expressions formulating a relationship between chemical structures and biological activities. These models have different capability, which is providing a deeper knowledge about the process of biological activity. In the first step of a usual QSAR study one needs to find a set of molecular descriptors with the higher influence on the biological activity of interest.[9] A broad scope of molecular descriptors [10] has been used in QSAR modeling. These molecular descriptors[11] have been categorised into different classes, including constitutional, geometrical, topological, quantum chemical and so on. Using such an way one could predict the activities of newly formulated compounds before a conclusion is being made whether these compounds should be truly synthesized and tested. We examine the performance of Partial Least Squares(PLS) based QSAR models with the results produced by Multi Linear Regression(MLR) and Principal Component Regression (PCR) methods to discover basic requirements for additional bettered antitubercular activity.

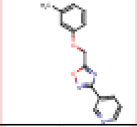
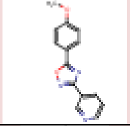
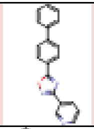
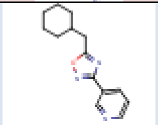
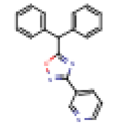
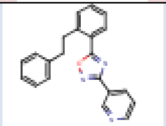
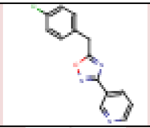
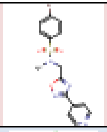
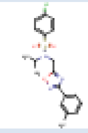
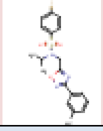
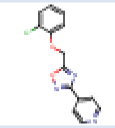
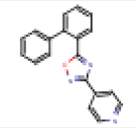
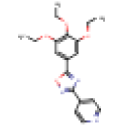
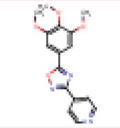
2. MATERIALS AND METHODS

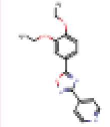
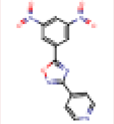
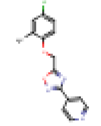
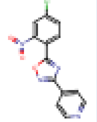
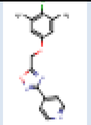
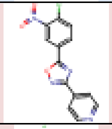
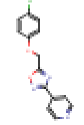
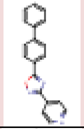
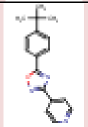
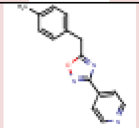
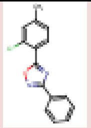
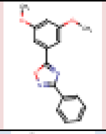
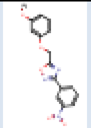
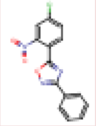
2.1 Molecular Descriptors Data Sets

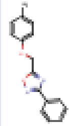
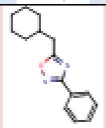
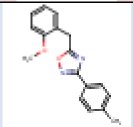
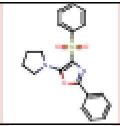
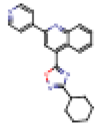
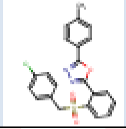
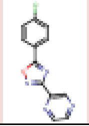
A set of fourty molecular compound structures relates to derivatives for mycobacterium TB (H37Rv) inhibitors were taken from large antitubercular drug molecular databases[12] using substructure mining tool Schrodinger Canvas 2010(Trial version).[13] All molecular structures were handled by the Vlife MDS[14] - 2D coordinates of atoms were recalculated counter ions and salts were eliminated from molecular structures, molecules were neutralized, mesomerized and aromatized. Data sets were then refined from duplicates. The 2D-QSAR models were produced using a training set of thirty molecular structures. Predictive ability of the models was assessed by a test set of ten molecular structures with consistently distributed biological activities. The observed selection of test set molecules was made by seeing the fact that test set molecules shows a range of biological activity similar to the training set. The actual and predicted biological activities of the training and test set molecular structures are given in Table 1.

Table 1 Molecular structures with Observed and Predicted activities of Nitrotriazole- and imidazole-based amides derivatives used in training and test set using MLR Expt. = Experimental activity, Pred. = Predicted activity IC50a = Compound concentration in micro mole required to inhibit growth by 50% PIC50b = -Log (IC50 × 10^{-6}): Training data set developed using MLR and Test set is light blue shaded.

Sl no	Compound	IC50a(µg/ ml)	PIC50b		Residual
			Expt	Pred	
1		3.85	5.414	4.8983	0.5157
2		2.33	5.632	5.5646	0.0674
3		6.42	5.192	5.3039	0.1119
4		1.04	5.982	5.8302	0.1518
5		0.84	6.075	5.2952	0.7798

6		3.6	5.443	5.4203	0.0227
7		7.07	5.150	5.6337	0.4837
8		26.23	4.581	5.0853	0.5043
9		9.55	5.019	5.5208	0.5018
10		20.8	4.681	4.6665	0.0145
11		6.13	5.212	3.9082	1.3038
12		2.72	5.565	5.2837	0.2813
13		0.49	6.309	5.9803	0.3287
14		1.49	5.841	4.0941	1.7469
15		23.57	4.271	4.4061	0.1351
16		12.92	4.888	5.3893	0.5013
17		13.09	4.883	5.0947	0.2117
18		4.03	5.394	4.7397	0.6543
19		1.41	5.850	5.7958	0.0542

20		12.67	4.897	5.1551	0.2581
21		97.9	4.009	4.2562	0.2472
22		9.04	5.043	5.2044	0.1614
23		3.73	5.428	5.2694	0.1586
24		5.72	5.242	4.9341	0.3079
25		0.83	6.083	5.4900	0.593
26		12.31	4.909	5.3703	0.4613
27		9.67	5.014	5.1325	0.1185
28		20.77	4.682	4.6617	0.0203
29		3.67	5.435	5.4565	0.0215
30		14.59	4.835	4.8422	0.0072
31		5.48	5.261	5.4146	0.1536
32		18.23	4.739	4.4836	0.2554
33		18.00	4.744	4.9716	0.2276

34		2.05	5.688	5.1024	0.5856
35		1.83	5.732	5.1711	0.5609
36		2.07	5.684	5.0541	0.6299
37		0.09	7.045	7.045	0.0000
38		99.29	4.003	3.5876	0.4154
39		6.51	5.186	3.2125	1.9735
40		0.31	6.508	6.4308	0.0772

2.2 Biological observed activities data

For the evolution of QSAR models of, Nitrotriazole- and imidazole-based amides of processes antitubercular activities in terms of half maximum inhibitory concentration IC50 (µM) versus (H37Rv) strains were took from the antitubercular drug molecular databases.[12] The IC50 activities data contains only molecular structures that have at least exhibited some activities. The biological activities data (IC50) were transformed in to pIC50 according to the formula pIC50 = ($-\log(\mathrm{IC50} \times 10^{-6})$) was used as response values, thus correlating the data linear to the free energy change.

2.3 Descriptor calculation for molecular dataset

The VLife MDS tool used for the computation of various molecular descriptors containing topological index (J), connectivity index (x), radius of gyration (RG), moment of inertia, Wiener index(W), balabian index(J), centric index, hosoya index (Z), information based indices, XlogP, logP , hydrophobicity, elemental count, path count, chain count, pathcluster count, molecular connectivity index (chi), kappa values, electro topological state indices, electrostatic surface properties, dipole moment, polar surface area(PSA), alignment independent descriptor (AI).[11,14] The calculated molecular descriptors were gathered in a data matrix. The preprocessing for the generated molecular descriptors was done by removing invariable (constant column) and cross-correlated descriptors (with r = 0.99). which happen in total 156, 125 and 162 molecular descriptors for MLR, PCR and PLS accordingly to be used for QSAR analysis.

2.4 Creation of Training and Test Set

The dataset of forty molecular descriptors is split s into training and test set by Sphere Exclusion (SE)[15-16] technique. In this technique initially data set splits into training and test set using sphere exclusion technique. In this technique variance value provides an idea to handle training and test set size. It needs to be adapted by trial and error until a desired split of training and test set is acquired. Increase in dissimilarity value results in increase in number of molecules in the test set. This technique is used for MLR, PCR and PLS models with pIC50 activity data as response variable and various 2D molecular descriptors computed for the molecules as independent variables.

2.5 Model Validation

Model validation [17-18] is a essential manner of quantitative structure–activity relationship (QSAR) modelling. This is done to test the internal stability and predictive capability of the QSAR models. These three QSAR models were validated by the following method.

2.5.1 Internal Model Validation

Internal model validation was carried out using leave-one-out (LOO-Q^2) method. For calculating q^2, each sample in the training set was eliminated once and the activity of the eliminated sample was predicted by using the model developed by the remaining samples. The Q^2 computed using the expression which explains the internal strength of a model.

$$Q^2 = 1 - \frac{\sum(Y_{pred} - Y_{obs})^2}{\sum(Y_{obs} - Y_{mean})^2} \qquad (1)$$

In Eq. (1), Y_{pred} and Y_{obs} indicate predicted and observed activity values accordingly and Y_{mean} signify mean activity value. A model is considered acceptable when the value of Q^2 exceeds 0.5.

2.5.2 External Model Validation

External model validation, the activity of each sample in the test set was predicted using the model created by the training set. The $\mathrm{pred\_r}^2$ value is computed as follows.

$$\mathrm{pred\_r}^2 = \frac{\Sigma(\mathrm{Y}_{\mathrm{pred(test)}} - \mathrm{Y}_{\mathrm{test}})^2}{\Sigma(\mathrm{Y}_{\mathrm{train}} - \mathrm{Y}_{\mathrm{mean(train)}})^2} \qquad (2)$$

In Eq (2) $\mathrm{Y}_{\mathrm{pred(test)}}$ and $\mathrm{Y}_{\mathrm{test}}$ indicate predicted and observed activity values for the test set and $\mathrm{Y}_{\mathrm{train}}$ indicates mean activity value of the training set. For the predictive QSAR model, the value of $\mathrm{pred\_r}^2$ must be more than 0.5.

2.5.3 Randomization Test

Randomization test or Y-scrambling is key mean of statistical validation. To assess the statistical importance of the QSAR model for the dataset, one tail hypothesis testing is used. The strength of the models for training sets was tested by examining these models to those derived for random datasets. Random sets were produced by rearranging the activities of the samples in the training set. The statistical model was determined using different randomly reorganize activities (random sets) with the chosen molecular descriptors and the equivalent Q^2 were computed. The importance of the models for that reason obtained was developed based on a computed $\mathrm{Z}_{\mathrm{score}}$.

A Z score value is calculated by the following equation:

$$\mathrm{Z}_{\mathrm{score}} = \frac{(h - \mu)}{\sigma} \qquad (3)$$

Where h is the Q^2 value computed for the dataset, μ the mean Q^2, and is its σ standard deviation calculated for various iterations using models build by different random datasets. The probability (a) of importance of randomization test is derived by comparing $\mathrm{Z}_{\mathrm{score}}$ value with $\mathrm{Z}_{\mathrm{score}}$ critical value as stated, if $\mathrm{Z}_{\mathrm{score}}$ value is less than 4.0; otherwise it is computed by the expression as given in the literature. For example, a $\mathrm{Z}_{\mathrm{score}}$ value more than 3.10 proposes that there is a probability (a) of smaller than 0.001 that the QSAR model build for the dataset is random. The randomization test proposes that all the created models have a probability of less than 1% that the model is produced by chance.

2.6 Multiple Linear Regression (MLR) Analysis

MLR technique used for modelling linear relationship between a response variable Y (pIC50) and independent variables X (2D molecular descriptors). MLR is based on least squares technique: the model is fit such that sum-of-squares of differences of actual and a predicted values are minimized. MLR estimates the regression coefficients (r^2) by applying least squares fitting technique. The model builds a relationship in the form of a straight line (linear) that best estimates all the individual data points. In regression analysis, conditional mean of response variable (pIC50) Y depends on (molecular descriptors)X. MLR analysis add to this idea to include more than one independent variables. Regression expression takes the form.

$$Y = b_1x_1 + b_2x_2 + b_3x_3 + c \qquad (4)$$

where Y is a response variable, 'b's are regression coefficients for corresponding 'x's are molecular descriptors(independent variables), 'c' is a regression constant or intercept [19,25].

2.7 Principal Component Regression (PCR) Analysis

Principal Component Regression (PCR) is a regression technique that uses principal component analysis(PCA) when evaluating regression coefficients. PCR presents a technique for finding structure in datasets. Its object is to group correlated variables, replacing the earlier descriptors by new set called principal components (PCs). These PC's are uncorrelated and are developed as a simple linear aggregation of earlier variables. It moves the data into a new set of axes such that first few axes indicates most of the variations within the data. First PC (PC1) is expressed in the direction of maximum variance of the whole dataset. Second PC (PC2) is the direction that defines the maximum variance in orthogonal subspace to PC1. Consequent components are taken orthogonal to the particular formerly chosen and defines best of remaining variance, by locating the data on new set of axes, it can points major fundamental structures certainly. Value of each point, when moved to a given axis, is called the PC value. PCA chooses a new set of axes for the data. These are chosen in decreasing order of variance within the data. The aim of principal component PCR is the computation of values of a response variable on the basis of chosen PCs of independent variables.[21]

2.8 Partial Least Squares (PLS) Regression Analysis

PLS is a well known regression technique which can be used to correlate one or more response variable (Y) to various independent variables(X) . PLS relates a matrix Y of response variables to a matrix X of molecular descriptors. PLS is useful in conditions where the number of molecular descriptors(independent variables) exceeds the number of samples, when X data contain colinearties or when N is less than 5M, where N is number of samples and M is number of response variables. PLS builds orthogonal components using existing correlations between independent variables and corresponding outputs while also keeping most of the variance of independent variables. Major aim of PLS regression is to predict the activity (Y) from X and to define their common frame.[22,23] PLS is probably the least contrary of various multivariate extensions of MLR model. PLS is a technique for constructing predictive models when factors are many and highly collinear.

2.9 Evaluation of the QSAR Models

The created QSAR models are computed using the following statistical parameters: N (Number of samples in regression); K (Number of independent variables (molecular descriptors)); DF (Degree of freedom); optimum component (number of optimums); r^2 (the squared correlation coefficient); F test (Fischer's Value) for statistical importance; q^2 (cross-validated correlation coefficient); $pred\_r^2$ (r^2 for external test set); Z_{score} (Z score computed by the randomization test); $Best\_ran\_r^2$ (maximal r^2 value in the randomization test) ; $Best\_ran\_q^2$ (maximal q^2 value in the randomization test); α (statistical importance parameter obtained by the randomization test). The correlation coefficient r^2 is a respective standard of fit by the regression expression. It expressed the part of the variation in the observed data is analyzed by the regression. Despite, a QSAR models are examined to be predictive, if the following prerequistes are satisfied: $r^2 > 0.6$, $q^2 > 0.6$ and $pred\_r^2 > 0.5$.[24] The F-test indicates the ratio of variance described by the model and variance due to the error in the regression. High values of the F-test indicate that model is statistically meaningful. The reduced standard error of $pred\_r^2se$, q^2_se and r^2_se demonstrates actual value of the fitness of the model. The cross-correlation extent was set at 0.5.

3. RESULTS

Taining set of 30 and 10 of test set of Nitrotriazole- and imidazole-based amides having different substitution were employed.

3.1 Creation of QSAR Models

3.1.1 Partial Least Squares (PLS) Regression Analysis

The molecular descriptors were applied to PLS technique to develop QSAR models by using simulated anealing variable selection mode. PLS model is having following QSAR Eq.(5) with four molecular descriptors.

$$pIC50 = 22.0067(chi5chain) - 0.6140\ (slogp) + 1.0048(SddsN(nitro)E - index) - 0.0025(XKHydrophobicArea) + 5.7409$$

Table 2 Statistical parameters of PLS, MLR And PCR

Parameters	PLS	MLR	PCR
N	40	40	40
DF	26	22	26
r^2	0.7523	0.8340	0.4009
q^2	0.7367	0.8123	0.3874
F-test	22.9825	28.1025	17.4020
$best\_ran\_r^{2}$	0.51614	0.58508	0.28130
$best\_ran\_q^{2}$	0.49415	0.55268	0.36218
$Z_{score\_ran\_r}{}^{2}$	8.35129	4.54816	8.41637
$Z_{score\_ran\_q}{}^{2}$	4.11500	2.03211	3.78423
$\alpha\_ran\_r^{2}$	0.00001	0.00009	0.00020
$\alpha\_ran\_q^{2}$	0.00100	0.05000	0.00100
r^2_se	0.3672	0.3268	0.5711
q^2_se	0.5967	0.5378	0.7714
$pred\_r^2$	0.7189	0.7945	0.3796
$pred\_r^2se$	0.6156	0.5678	0.7652

The above analysis directs to the improvement of statistically meaningful QSAR model, which allows understanding of the molecular properties/features that play an key role in governing the variation in the activities. In addition, this QSAR study allowed examining influence of very simple and easy-to-compute molecular descriptors in discovering biological activities, which could shed light on the important factors that may aid in design of new potent molecules.

All the parameters and their significance, which contributed to the specific antitubercular inhibitory activity in the generated model is discussed below.

1. chi5chain: This descriptor signifies a retention index for five membered ring. Positive Contribution of this descriptor to the model is 35.62%.

2. slogp: This descriptor signifies the log of the octanol/water partition coefficient (including implicit hydrogen) which is a measure of the lipophilicty of the molecule. Negative Contribution of this descriptor to the model is -30.31%.

3. SddsN(nitro)E-index: This descriptor indicates the Electrotopological state indices for number of –nitro group connected with two double and one single bond. Positive Contributions of this descriptor to the model is 20.76%.

4. XKHydrophobicArea: This descriptor describes the vdW surface descriptor showing hydrophobic surface area. (By Kellog Method using Xlogp) . Negative Contribution of this descriptor to the model is --13.31%.

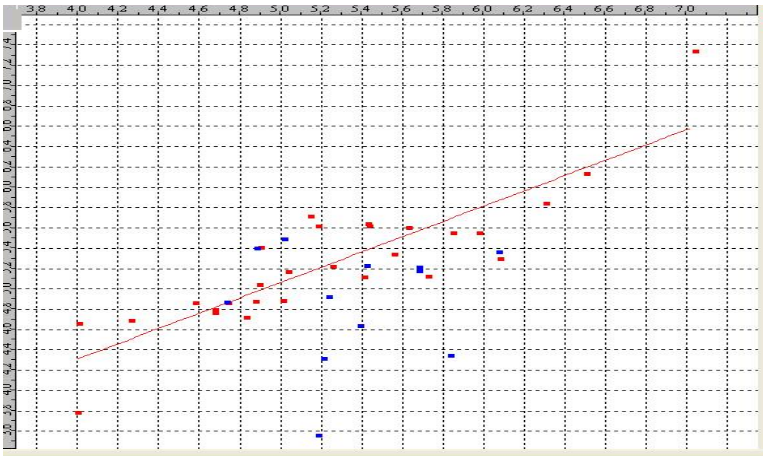


Figure 1: Observed vs. Predicted activities for training and test set molecular descriptors by Partial Least Square model. (A) Training set (Red dots) (B) Test Set (Blue dots).

The PLS model gave correlation coefficient (r^2) of 0.7523, significant cross validated correlation coefficient (q^2) of 0.7367, F-test of 28.1025 and degree of freedom 22. The model is validated by α_ran_r^2 = 0.00001, α_ran_q^2 = 0.00100, best_ran_r^2 = 0.51614, best_ran_q^2 = 0.49415, $Z_{score\_ran\_r}{}^2$= 4.54816 and $Z_{score\_ran\_q}{}^2$ = 4.11500. The randomization test proposes that the created model have a probability of smaller than 1% that the model is build by chance. Statistical data is presented in Table 2. The graph of observed vs. predicted activity is demonstrated in Figure 1. The descriptors which contribute for the QSAR model is demonstrated in Figure 2.

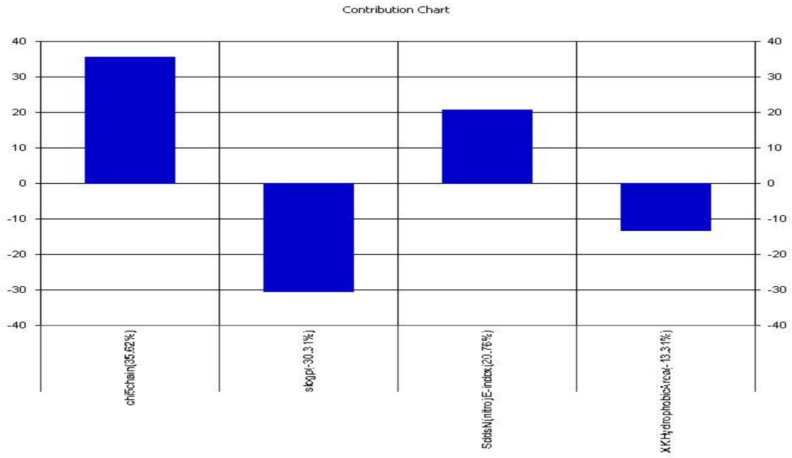


Figure 2: Percentage contribution of each molecular descriptor in created PLS model describing variation in the activities

3.1.2 Multiple Linear Regression (MLR) Analysis

The QSAR analysis by Multiple Linear Regression method with simulated annealing variable selection technique, the final QSAR model is created having five molecular descriptors is shown in Eq. (6).

$$\mathrm{pIC50} = 21.6174(\pm 2.7444)\mathrm{chi5chain} - 0.6014(\pm 0.0213)\mathrm{slogp} + 1.5239(\pm 0.1047)\mathrm{SddsN(nitro)E-index} - 0.0043(\pm 0.0000)\mathrm{XKHydrophobicArea} + 4.4022(\pm 1.7117)\ \mathrm{SAMostHydrophobic} + 5.2851 \quad (6)$$

MLR Model has a correlation coefficient (r^2) of 0.8340, significant cross validated correlation coefficient (q^2) of 0.8123, F test of and degree of freedom 24. The model is validated by $\alpha\_ran\_r^2$ = 0.00009, $\alpha\_ran\_q^2$= 0.05000, $best\_ran\_r^2$ = 0.58508, $best\_ran\_q^2$ = 0.55268 , $Z_{score\_ran\_r}{}^2$ = 8.11471 and $Z_{score\_ran\_q}{}^2$ = 2.03211 . The randomization test proposes that the created model have a probability of smaller than 1% that the model is build by chance. The observed and predicted values with residual values are demonstrated in Table 1.Statistical data is demonstrated in Table 2.The graph of observed vs. predicted activity demonstrated is in Figure 3. The descriptors which contribute for the QSAR model are demonstrated in Figure 4. All the parameters and their significance, which contributed to the specific antitubercular inhibitory activities in the generated models are explained below.

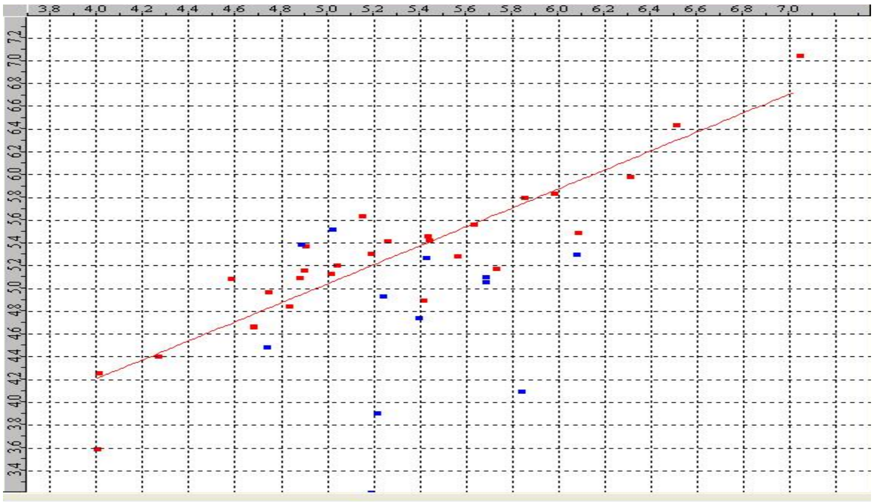


Figure 3 Observed vs. Predicted activities for training and test set molecular descriptors from the Multiple Linear Regression model. (A) Training set (Red dots) (B) Test Set (Blue dots).

1.chi5chain: This descriptor signifies a retention index for five membered ring. Positive Contribution of this descriptor to the model is 26.73%.

2. slogp: This descriptor signifies the log of the octanol/water partition coefficient (including implicit hydrogen) which is a measure of the lipophilicty of the molecule. Negative Contribution of this descriptor to the model is -22.68%.

3. SddsN(nitro)E-index: This descriptor indicates the Electrotopological state indices for number of –nitro group connected with two double and one single bond. Positive Contributions of this descriptor to the model is 24.05%.

4. XKHydrophobicArea: This descriptor describes the vdW surface descriptor showing hydrophobic surface area. (By Kellog Method using Xlogp). Negative Contribution of this descriptor to the model is --17.88%.

5. SAMostHydrophobic: This descriptor defines most hydrophobic value on the vdW surface. (By Audry Method using Slogp). Positive Contribution of this descriptor to the model is 8.66%.

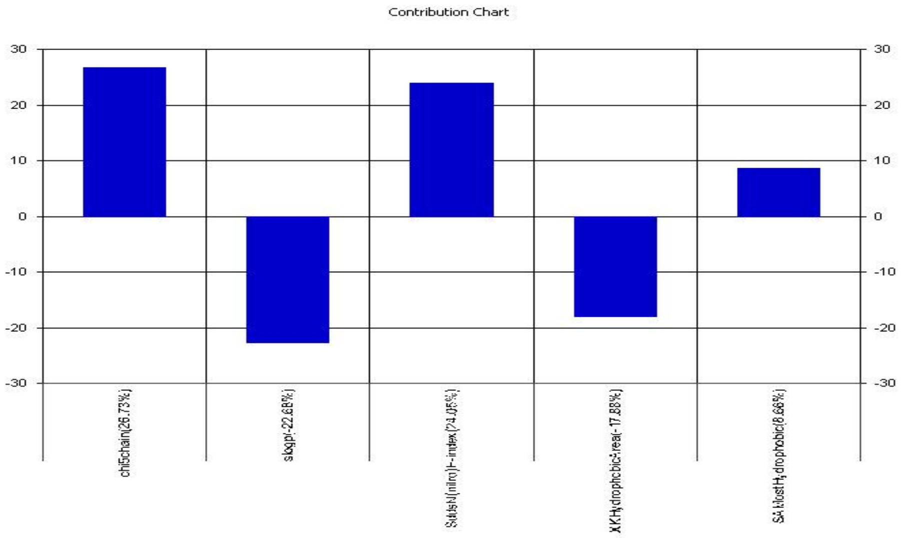


Figure 4: Percentage contribution of each molecular descriptor in created MLR model describing variation in the activities.

3.1.3 Principal Component Regression (PCR) Analysis

The molecular descriptors were applied to under goes PCR technique to create QSAR model with Simulated anealining variable selection mode by using PCR model. The final QSAR model is Eq. (7) was created having two molecular descriptors as follows.

$$\text{pIC50} = 13.4249\ (\text{chi5chain}) - 0.6243\ (\text{6ChainCount}) + 4.9970 \qquad (7)$$

The PCR model gave correlation coefficient (r^2) is 0.4009, significant cross validated correlation coefficient (q^2) of 0.3874, F test of 17.4020 and degree of freedom 26. The model is validated by $\alpha\_ran\_r^2$ = 0.00020, $\alpha\_ran\_q^{\ 2}$ = 0.00100, $best\_ran\_r^{\ 2}$ = 0.28130, $best\_ran\_q^2$ = 0.36218, $Z_{score\_ran\_r}{}^{2}$ = 8.41637 and $Z_{score\_ran\_q}{}^{2}$ =.3.78423 The randomization test proposes that the created model have a probability of smaller than 1% that the model is build by chance. Statistical data is demonstrated in Table 2. The graph of observed vs. predicted activity is in demonstrated Figure 5 .The descriptors which contribute for the QSAR model is demonstrated in Figure 6.

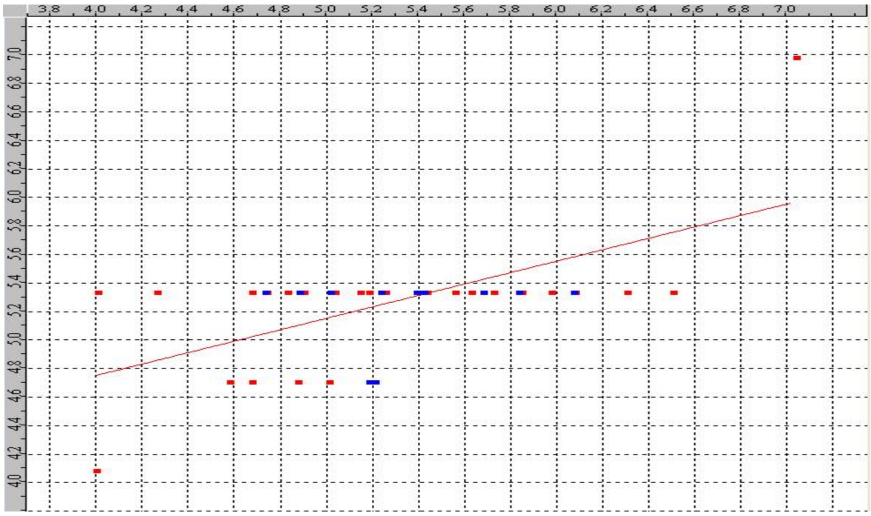


Figure 5: Observed vs. Predicted activities for training and test set molecular descriptors by Principal Component Regression model. A) Training set (Red dots) B) Test Set (Blue dots).

All the parameters and their significance, which contributed to the specific antitubercular inhibitory activity in the generated models are discussed here.

1.chi5chain: This descriptor signifies a retention index for five membered ring. Positive Contribution of this descriptor to the model is 50.00%.

2. 6ChainCount: This descriptor signifies total number six membered rings in a compound. Negative Contribution of this descriptor to the model is -50.00%.

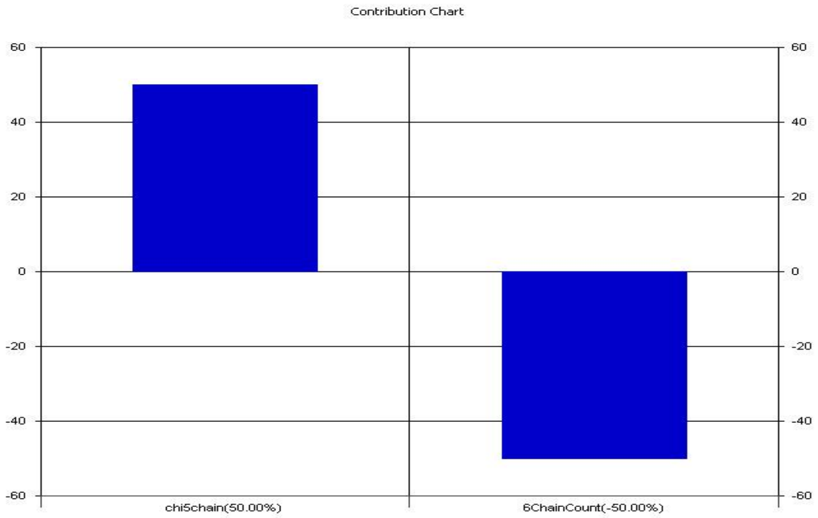


Figure 6 Percentage contribution of each molecular descriptor in developed PCR model describing variation in the activities

4. CONCLUSION

The 2D QSAR analysis were conducted with a series of Nitrotriazole- and imidazole-based amides derivatives for mycobacterium tuberculosis(H37Rv) inhibitors, and some useful predictive models were obtained. The physicochemical molecular descriptors were found to have an key role in governing the change in activity. The statistical parameters demonstrate the estimation power of QSAR model for the molecular descriptor data set from which it has been determined and evaluate it only internally. The overall performance of prediction was found to be around 84% in case of PLS and MLR. Among the three 2D-QSAR models (MLR, PCR, and PLS), the results of PLS and MLR analysis showed significant predictive power and reliability as compare to PCR technique.

ACKNOLDGEMENTS

The Authors are thankful to Department of Pharmaceutical Chemistry K.L.E Pharmacy College Hubli.

REFERENCES

1. “Tuberculosis”, Centers for Disease Control and Prevention1600 Clifton Rd. Atlanta, GA 30333, USA http://www.cdc.gov/tb/topic/basics/default.htm.
2. “Weekly Epidemiological Record (WER)”,WHO annual report on global TB control – summary http://www.who.int/wer/2003/wer7815/en/index.html

3. Phyllis C. Braun, PhD and John D. Zoidis, MD. "Update on Drug-Resistant Pathogens: Mechanisms of Resistance, Emerging Strains", http://www.rtmagazine.com/issues/articles/2004-01_01.asp
4. "Tuberculosis management", From Wikipedia, the free encyclopaedia http://en.wikipedia.org/wiki/Tuberculosis_management
5. "Multidrug-resistant tuberculosis (MDR-TB)", From World Health Organization http://www.who.int/tb/challenges/mdr/en/
6. Papadopoulou MV, Bloomer WD, Rosenzweig HS, Arena A, Arrieta F, Rebolledo JC, Smith DK.(2014),"Nitrotriazole- and imidazole-based amides and sulfonamides as antitubercular agents derivatives", Antimicrob Agents Chemother. 2014 Nov; 58(11): 6828-36. doi: 10.1128/AAC.03644-14. Epub 2014 Sep 2.
7. Ujjini H. Manjunatha, Helena Boshoff, Cynthia S. Dowd, Liang Zhang, Thomas J. Albert, Jason E. Norton, Lacy Daniels, Thomas Dick, Siew Siew Pang, and Clifton E. Barry, "Identification of a nitroimidazo- oxazine-specific protein involved in PA-824 resistance in Mycobacterium tuberculosis", Proceedings of the National Academy of the Sciences, vol. 103 no. 2,431–436, doi: 10.1073/pnas.0508392103
8. R. Karbakhsh1,* and R. Sabet (2011),"Application of different chemometric tools in QSAR study of azoloadamantanes against influenza A virus", Research in Pharmaceutical Sciences; 6(1): 23-33
9. "Molecular Descriptors", The Free Online resource, http://www.moleculardescriptors.eu/tutorials/what_is.htm
10. "Molecular Descriptors Guide", Version 1.0.2 Copyright [2008] US Environamental Protection agency.
11. "Streamline Drug Discovery with CDD colabrative web based software", https://www.collaborativedrug.com/ (Accesed in May-june [2012])
12. "Canv as", A comprehensive cheminformatics computing environment http://www.schrodinger.com/products/14/23/
13. "VlifeMDS", Integrated platform for Computer Aided Drug Design (CADD) http://www.vlifesciences.com/products/VLifeMDS/Product_VLifeMDS.php
14. "Sphere Exclusion Method for set selection", Rajarshi Guha Penn State University http://rguha.net/writing/pres/tropsha.pdf
15. Mary Ann Liebert. "Dissimilarity-Based Algorithms for Selecting Structurally Diverse Sets of Compounds". Journal of Computational Biology, 1991; 6(3/4): Inc. Pp. 447–457.

16. Tropsha, A.; Gramatica, P.; Gombar, V.K.(2003).” The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models.” QSAR Comb. Sci., 22, 69-77.
17. Partha Pratim Roy, Somnath Paul, Indrani Mitra and Kunal Roy(2009),“On Two Novel Parameters for Validation of Predictive QSAR Models”, Molecules, 14, 1660-1701 ISSN 1420-3049.
18. “Multile linear Regression”, http://www.ltrr.arizona.edu/~dmeko/notes_11.pdf
19. Dr. Frank Dieterle,“Variable Selection by Simulated Annealing”, http://www.frank-dieterle.de/phd/2_8_6.html
20. Hwang , Dan Nettleton,“Principal Components Regression With Data-Choosen Components and related methods”, J.T. Gene www.math.cornell.edu/~hwang/pcr.pdf
21. Herv´e Abdi1, “Partial Least Squares(PLS) Regression.” The University of Texas at Dallas.
22. Randall D. Tobias, “An introduction to partial least squares Regression”, SAS Institute Inc., Carry, NC www.ats.ucla.edu/stat/sas/library/pls.pdf
23. Golbraikh. A, and A. Tropsha, (2002),”Predictive QSAR modeing based on diversity of sampling of experimental datasets for the training and test set selection“, J. Comp Aided. Mol Design, 16: 357-366.
24. “Influence of observations on the misclassification probability in quadratic discriminant analysis”. https://lirias.kuleuven.be/bitstream/123456789/85608/1/qda.pdf
25. Craig A. James, “An introduction to the Computer Science and Chemistry of Chemical Information Systems”, eMolecules, Inc. http://www.emolecules.com/doc/cheminformatics-101.php