Available on line www.jocpr.com

Journal of Chemical and Pharmaceutical Research

J. Chem. Pharm. Res., 2010, 2(2): 357-365

ISSN No: 0975-7384

Quantitative structure-activity relationship analysis of some 2-substituted halogenbenzimidazoles analogues using computer-aided drug designing technique

Smita Sharma *[a], Mukesh Chandra Sharma [b] and A. D. Sharma[c]

[a] *Department of Chemistry, Yadhunath Mahavidyalya, Bhind (M.P), India*
*[b] *School of Pharmacy, Devi Ahilya Vishwavidyalaya, Indore (M.P), India*
[c] *Oriental College of Pharmacy, Indore(M.P), India*

Abstract

Tuberculosis, which is caused by single infectious agent Mycobacterium tuberculosis, is one of the most important infectious diseases. Tuberculosis is a major public health problem with approximately 2 million annual deaths. In the present study QSAR analysis of a series of substituted 2-polyfluoroalkyl and 2-Nitrobenzylsulphanyl benzimidazoles was performed using V-LIFE MDS 3.0 software 2D QSAR models were developed using partial least square (PLS) and variable selection methods. Out of 10 models developed. the two best 2D QSAR models having highest correlation coefficient and cross validated squared correlation coefficient were selected for further study, which were r2 = 0.9013, q2 = 0.7676, F test = 73.0375 pred_r2 = -0.1772, pred_r2se = 0.5646 and r2 = 0.8441, q2 = 0.7088, F test = 27.0819 pred_r2 = -0.1082, pred_r2se = 0.5478 . Two 3D QSAR models were developed using KNN-MFA method, combined with simulated annealing selection procedure. Out of two models developed the best 3D QSAR model having highest cross validated squared correlation coefficient was selected for further study, which is q2 = 0.6765 pred_r2se = 0.5312 and q2 = 0.7747, pred_r2se = 0.8455 . A quantitative structure activity relationship study on a series of Halogenbenzimidazoles analogues was made using combination of various thermodynamic electronic and spatial descriptors. Several statistical expressions were developed using stepwise multiple liner regression analysis. The best quantitative structure activity relationship models were further validated by leave-one-out method of cross-validation.

Key words: 2D QSAR, PLS Regression, Antimycobacterial, Halogenbenzimidazoles.

Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, remains the leading cause of mortality due to a bacterial pathogen. WHO estimated that there were 8.8 million new cases of tuberculosis in 2020. No new drug against tuberculosis has been developed in the last 30 years. The main objective of the present study was the search for novel benzimidazole compounds that would show a promise to become useful antimycobacterial agent[1-4]. A series of compounds of 2-polyfluoroalkyl and 2-Nitrobenzylsulphanyl benzimidazoles was selected as novel antimycobacterial agents for QSAR studies. Tuberculosis (TB) is one of the oldest and pervasive diseases in history caused by respiratory infection, by a gram-positive bacteria Mycobacterium tuberculosis[5]. In recent years, TB has re-emerged as a major world health problem with an estimated annual death toll of 2 million. That's why TB remains a major world health problem. There were several drugs discovered for the treatment of TB since 1940 but due to drug resistance cases there is a continuing need to find additional lead compounds and biological targets for novel anti tubercular chemotherapies[6]. As napthoquinolones were widely distributed in plants, fungi and some animals and many are found to exhibit various pharmacological actions like antibacterial, antimalarial, antiviral, trypanocidal, anticancer and antifungal activity[7-14].

Materials and Methods

Experimental

The Data sets The data set used for the QSAR analyses contains 28, 2-polyfluoroalkyl and 2-Nitrobenzylsulphanyl benzimidazoles was selected as novel antimycobacterial agents for QSAR studies. All the structures of the compounds were drawn in 2D-APPL mode of software and exported to 3D model. The chemical structure and their corresponding IC_{50} values were mentioned in Table I. The modeling analyses, calculations, and visualizations for 2D QSAR were performed using the V-Life Molecular Design Suite 3.0 (Vlife MDS) [15]. A set of 28 molecules was selected and divided in training (19) and test set (9).The negative logarithm of IC_{50} values (PIC_{50}) calculated using the IC_{50} values of reported compounds. The biological activity data (IC_{50} in Molar) were converted in to pIC_{50} according to the formula $pIC_{50} = (-\log (IC_{50})$. Thus such studies may help for the design and synthesis of better 2-polyfluoroalkyl and 2-Nitrobenzylsulphanyl benzimidazoles. All the twenty eight compounds were built on workspace of molecular modeling software V-Life MDS 3.5, which is a product VLife Sciences Pvt Ltd., India. The compounds were then subjected to conformational analysis and energy minimization using montocarlo conformational search with RMS gradient of 0.001 kcal/mol and iteration limit of 10000 using a MMFF94 force field. Montocarlo conformational search method is similar to the RIPS method that generates a new molecular conformation by randomly perturbing the position of each coordinate of each atom in molecule, followed by energy minimization and optimization is necessary process for proper alignment of molecules around template. Most stable structure for each compound was generated after energy minimization and used for calculating various physico-chemical descriptors like thermodynamic, steric and electronic. The various descriptors selected for 2D QSAR were vdWSurfaceArea (van der Waals surface area of the molecule), –vePotential Surface Area (total van der Waals surface area with negative electrostatic potential of the molecule), +vePotentialSurfaceArea (total van der Waals surface area with positive electrostatic potential of the molecule) dipole moment, YcompDipole (y component of the dipole moment), element count, slogP, path count, cluster, distance based topological indices, connectivity index, hydrophobic and hydrophilic areas like SA Most Hydrophilic (Most hydrophilic value on the vdW surface by Audry Method using Slogp), SAMostHydrophobicHydrophilic Distance (distance between most hydrophobic and hydrophilic point on the vdW surface by Audry Method using Slogp), SAHydrophilicArea (vdW surface

descriptor showing hydrophilic surface area by Audry Method using SlogP) and SKMostHydrophilic (Most hydrophilic value on the vdW surface by Kellog Method using Slogp), radius of gyration, Wiener's index, moment of inertia, semi- empirical descriptors, HOMO (Highest occupied molecular orbital), LUMO (lowest unoccupied molecular orbital), heat of formation and ionization potential. Besides these all alignment independent descriptors were also calculated. The hydrophobic descriptors govern the movement of a drug molecule across the biological membranes in order to interact with the receptor by vander Waals binding forces whereas both electronic and steric descriptors influence the affinity of a drug molecule necessary for proper drug- receptor interaction. The optimal training and test sets were generated by either random selection method or the sphere exclusion algorithm. A commonly used ratio of training to validation objects (test set), which was also adopted in this work, is 70%: 30% [9]. However, rational splitting was accomplished by applying a sphere-exclusion type algorithm [10-14]. In classical sphere-exclusion algorithm the molecules are selected whose similarities with each of the other selected molecules are not higher than a defined threshold. Each selected molecule generates a hyper-sphere around itself, so that any molecule inside the sphere is excluded from the selection in the train set and driven toward the test set. The number of compounds selected and the diversity among them can be determined by adjusting the radius of the sphere (R). All the molecules were optimized (energy minimization) MMFF using the software V_LIFE MDS 3.0. various 2D descriptors like, T_C_O_6, T_Cl_Cl_3, T_T_C_5, T_N_F_4, T_N_N_6, T_2_Cl_5, T_T_T_7, SsBrE- index count that are responsible for antimycobacterial activity were calculated. The different statistical models were developed using partial least square (PLS) method. The showed the better correlation between biological activity and physicochemical descriptor values. The correlation coefficient (r^2 value) was found 0.9013 and cross validated squared correlation coefficient value(Q^2) was found 0.6765 ,the other relevant data was found F test =73.0375, pred_r2 = -0.1772 , pred_r2se = 0.5646 (for model 1)and correlation coefficient (r2) was found 0.8441 , cross validated squared correlation coefficient value (Q2)was found 0.6765,and the other value were found to be F test = 20.2339 r2 se = 0.2438 q2 se = 0.2660 pred_r2 = -0.1045 pred_r2se = 0.5469(for model 2).The equation were generated for assuming the biological activity with the help of physicochemical descriptor values. The equation showed the correlation between biological activity and physioco- chemical descriptor values. The equation were found to be to derive 2D-QSAR equation different model building method (multiple regression, principle component regression) coupled with stepwise variable selection was used. Then QSAR models were generated by using partial linear regression method (PLS) method, by setting cross correlation limit as 0.5, number of variable in final equation as 5, and term selection criteria as r^2, F-test 'in' as 4 and F-test 'out' as 3.99. Variance cut off was set to 0 and scaling as auto scaling, number of random iteration was set to 10. Following statistical parameters were considered to compare the generated QSAR models: correlation coefficient (r), squared correlation coefficient (r^2), predicted r^2 (pred_ r^2), and Fischer's value (F). In order to validate the generated QSAR models Leave One out (LOO) method was used indicated as value of q^2 (cross- validated explained variance) which is a measure of internal predictive ability of the model.

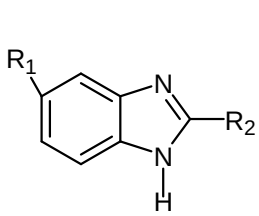
Statistical analysis

Models were generated by using three significant statistical methods, namely, partial least square analysis, multiple regressions, and principle component analysis. The cross-validation analysis was performed using the leave-one-out method. In the selected equations, the cross-correlation limit was set at 0.5, the number of variables at 10, and the term selection criteria at r^2. An F value was specified to evaluate the significance of a variable. The higher the F value, the more stringent was the significance level: F test ''in'' as 4 and F test ''out'' as 3.99. The variance cutoff was set at 0, and scaling was auto scaling in which the number of random iterations was

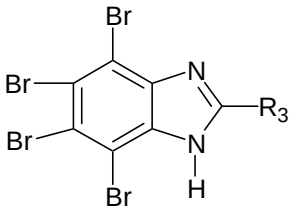
set at 100.The following statistical parameters were considered for comparison of the generated QSAR models: correlation coefficient (r), squared correlation coefficient (r2), predictive r2 for external test set (pred r2) for external validation, and Fischer's (F).The predicted r2 (pred_r2) value was calculated using Eq. 1, where yi and yˆi are the actual and predicted activities of the i^{th} molecule in the test set, respectively, and ymean is the average activity of all molecules in the training set. Both summations are over all molecules in the test set. The pred_r^2 value indicates the predictive power of the current model for the external test set as follows

$$\text{pred\_}r^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{mean})^2} \qquad (1)$$

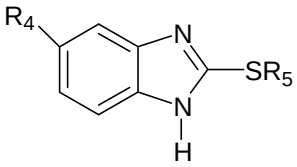
Table I- series of compounds of 2- Substituted Halogenobenzimidazoles with I_{C50} and P_{IC50} values



2a-2l



3a-3d



5a-5l

S.No	Compound	R^1	R^2	IC $_{50}$	Log IC $_{50}$
1.	2a	$4,6Cl_2$	Cf_3	32	1.505150
2.	2b	4,6 Cl_2	C_2f_5	16	1.204120
3.	2c	4,6 Cl_2	C_3f_7	8	0.903090
4.	2d	4,6 Cl_2	C_4f_9	8	0.903090
5.	2e	5,6 Cl_2	Cf_3	8	0.903090
6.	2f	5,6 Cl_2	C_2f_5	4	0.602060
7.	2g	5,6 Cl_2	C_3f	8	0.903090
8.	2h	5,6 Cl_2	C_4f_9	4	0.602060
9.	2i	$4,6Br_2$	Cf_3	32	1.505150
10.	2j	$4,6Br_2$	C_2f_5	16	1.204120
11.	2k	$4,6Br_2$	C_3f_7	16	1.204120
12.	2l	$4,6Br_2$	C_4f_9	16	1.204120

		R^3	IC $_{50}$	Log IC $_{50}$
13	3a	Cf_3	32	1.505150
14	3b	C_2f_5	8	0.903090
15	3c	C_3f_7	4	0.602060
16	3d	C_4f_9	16	1.204120

Internal validation was carried out using leave-one-out (q2, LOO) method. For calculating q2, each molecule in the training set was eliminated once and the activity of the eliminated molecule was predicted by using the model developed by the remaining molecules. The q2 was calculated using the equation which describes the internal stability of a model:

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{mean})^2} \quad (2)$$

Where y_i and $\hat{y}_i$ are the actual and predicted activity of the $_i$th molecule in the training set, respectively, and y_{mean} is the average activity of all molecules in the training set.

		R^4	R^5	IC_{50}	Log IC_{50}
17	5a	5-cl	3,5-dinitrobenzyl	2	0.301030
18	5b	5-Br	3,5-dinitrobenzyl	4	0.301030
19	5c	5-I	3,5-dinitrobenzyl	2	0.602060
20	5d	4,6- Cl_2	4-nitrobenzyl	2	0.301030
21	5e	4,6- Cl_2	2,4-dinitrobenzyl	32	1.505150
22	5f	4,6- Cl_2	3,5-dinitrobenzyl	4	0.602060
23	5g	4,6-Br_2	4-nitrobenzyl	2	0.301030
24	5h	4,6-Br_2	2,4-dinitrobenzyl	16	0.301030
25	5i	4,6-Br_2	3,5-dinitrobenzyl	4	0.602060
26	5j	4,5,6,7-Br_4	4-nitrobenzyl	16	0.602060
27	5k	4,5,6,7-Br_4	2,4-dinitrobenzyl	16	1.204120
28	5l	4,5,6,7-Br_4	3,5-dinitrobenzyl	8	0.903090

Table II- Calculated descriptors for 2D QSAR

No.	Compound	T_C_O_6	T_T_C_5	T_N_N_6	T_2_Cl_5	T_T_T_7
1.	2a	0	13	0	1	3
2.	2b	0	21	0	1	12
3.	2c	0	29	0	1	18
4.	2d	0	34	0	1	24
5.	2e	0	13	0	2	3
6.	2f	0	21	0	2	12
7.	2g	0	29	0	2	18
8.	2h	0	34	0	2	24
9.	2i	0	13	0	0	3
10.	2j	0	21	0	0	12
11.	2k	0	29	0	0	18
12.	2l	0	34	0	0	24
13.	3a	0	14	0	0	6
14.	3b	0	22	0	0	18
15.	3c	0	24	0	0	24
16.	3d	0	29	0	0	30
17.	5a	4	27	4	2	23
18.	5b	4	27	4	0	23
19.	5c	4	27	4	0	23
20.	5d	0	23	0	1	17
21.	5e	0	23	0	1	17
22.	5f	4	27	4	1	25
23.	5g	0	23	0	0	17
24.	5h	0	27	0	0	23
25.	5i	4	27	4	0	25
26.	5j	0	25	0	0	20
27.	5k	0	29	0	0	27
28.	5l	4	29	4	0	28

Table III - Actual and predicted activities of training and test set compounds in statistically significant models

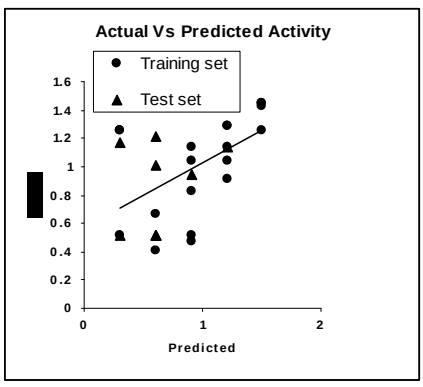
S.No.	Molecule	Actual Activity	Predicted Activity-1	Predicted Activity-2
1.	2a	1.505150	1.450495	1.450495
2.	2b	1.204120	1.292837	1.292837
3.	2c	0.903090	1.135178	1.135178
4.	2d	0.903090	1.036642	1.036642
5.	2e	0.903090	0.826724	0.826724
6.	2f	0.602060	0.669066	0.669066
7.	2g	0.903090	0.511407	0.511407
8.	2h	0.602060	0.412871	0.412871
9.	2i	1.505150	1.450495	1.450495
10.	2j	1.204120	1.292837	1.292837
11.	2k	1.204120	1.135178	1.135178
12.	2l	1.204120	1.036642	1.036642
13.	3a	1.505150	1.430788	1.430788
14.	3b	0.903090	0.943552	0.943552
15.	3c	0.602060	1.013997	1.013997
16.	3d	1.204120	0.91546	0.91546
17.	5a	0.301030	0.510417	0.510417
18.	5b	0.301030	0.510417	0.510417
19.	5c	0.602060	0.510417	0.510417
20.	5d	0.301030	1.253422	1.253422
21.	5e	1.505150	1.253422	1.253422
22.	5f	0.602060	0.510417	0.510417
23.	5g	0.301030	1.253422	1.253422
24.	5h	0.301030	1.174593	1.174593
25.	5i	0.602060	0.510417	0.510417
26.	5j	0.602060	1.214008	1.214008
27.	5k	1.204120	1.135178	1.135178
28.	5l	0.903090	0.471002	0.471002

Table IV Unicolumn statistics of Training and Test sets

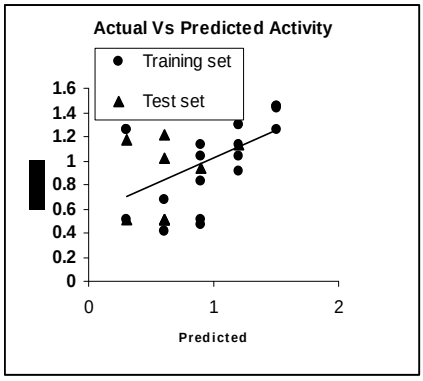
Activity (pic50)	Average	Maximum	Minimum	Std. Dev	Sum
Training set	0.8562	2.1744	0.0792	0.5305	11.9872
Test set	1.0857	1.4440	0.7993	0.2675	6.5140

Figure- Actual and Predicted values for model-1 and model-2, model 3 for 2D QSAR analysis R^2= 0.7961 (model-1) R^2= 0. 7152(model-2) R^2= 0. 7980 (model-3)

Model-I



Model-II



Model-III

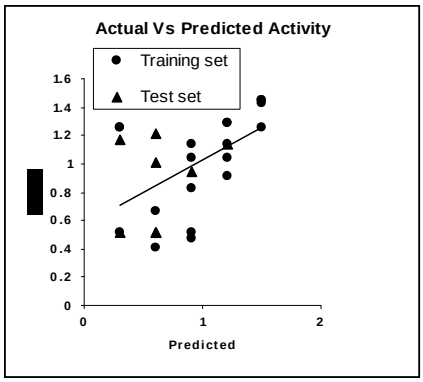


Table V- Actual Activity, Predicted Activity and Residual values of test set Compounds

S.No.	Molecule	Actual Activity	Predicted Activity-1	Predicted Activity-2
1	5d	0.301030	1.253422	1.253422
2.	5e	1.505150	1.253422	1.253422
3.	5f	0.602060	0.510417	0.510417
4.	5g	0.301030	1.253422	1.253422
5.	5h	0.301030	1.174593	1.174593
6	5i	0.602060	0.510417	0.510417
7.	5j	0.602060	1.214008	1.214008
8.	5k	1.204120	1.135178	1.135178
9	5l	0.903090	0.471002	0.471002

Result and Discussion

Biological activity data and various physico-chemical parameters were taken as dependent and independent variables and correlations were established using PLS method. When the compounds were subjected to under goes PLS method to developed QSAR models by using step wise forward-backward variable selection mode, four QSAR models, Model-I and Model-II, Model-III were developed for both the methods respectively as shown below and other good model predicted activity shown abstract.

Log_{10}(IC_50) = 2.431 T_O_O_2 - 4.1233 Hydrogen count - 0.0197 T_T_O_2 - 0.0549 T_N_N_4 + 1.7067 **(Model 1)**

Optimum Components = 2, Degrees of Freedom = 12, n = 28, r^2= 0.7961, q2= 0.5916, F test = 34.374 r2 se = 0.441, q2 se = 0.4690, pred_r^2 = 0.6961, SEE = 0.141, SECV= 0.210, SEP=0.290,

best_ran_r^2=0.065, best_ran_q^2= 0.116 Zscore_ran_r2 =0.083, Zscore_ran_q2= 0.102, α_ran_r^2 =<0.0001, α _ran_q^2 = <0.001

To improve the external predictivity of the model, PLS analysis with the same data set was performed, which resulted in a coefficient of correlation of 0.4462 and an internal predictive power of 31%, with the good external predictivity of 69.9%. Hydrogen count contributes in the same manner as above. T_O_O_2 defines the total number of carbons connected with four single bonds and makes a negative contribution to activity.

Log_{10}(IC_50) = + 2.6838 slogp + 0.6315 T_Cl_Cl_6 + 0.8013 Cluster +0.5364 +0.694 T_N_N_5 - 0.7845 T_N_N_5 **(Model 2)**

Optimum Components = 2, Degrees of Freedom = 12, n = 28, r^2= 0.7152, q2= 0.631, F test = 38.431, r2 se = 0.2351, q2 se = 0.6410, pred_r^2 = 0.7512, SEE = 0.223, SECV= 0.114, SEP=0.090, best_ran_r^2 = 0.365, best_ran_q^2 = 0.216, Zscore_ran_r2 =0.231, Zscore_ran_q2= 0.102, α_ran_r^2 =<0.0001, α _ran_q^2 = <0.001

Model –2 shows good squared correlation coefficient (r^2) of 0.7512 explains 73.12% variance in biological activity. This model also indicates statistical significance >99.9% with F values F = 38.431. Cross validated squared correlation coefficient of this model was 0.8039, which shows the good internal prediction power of this model. The graph of observed vs. predicted biological activities for the training and the test molecules is shown in Figure.

Log_{10} (IC_50) = + 0.8838 H-Donor Count + 0.6714 T_2_Cl_6 + 0.6073 chi5chain+0.5364 +0.394 T_N_N_5 - 0.3801 T_2_O_4 + 3.6338 **(Model 3)**

Optimum Components = 2, Degrees of Freedom = 12, n = 28, r^2= 0.7980 q2= 0.7673, F test 29.321 r2 se = 0.5251, q2 se = 0.6481, pred_r^2 = 0.8214, SEE = 0.091, SECV= 0.077, SEP=0.319, best_ran_r^2 = 0.212, best_ran_q^2 = 0.431, Zscore_ran_r2 =0.431, Zscore_ran_q2= 0.179, α_ran_r^2 =<0.00001, α _ran_q^2 = <0.01

Model – 3 shows good squared correlation coefficient (r^2) of 0.7980 explains 79.80 % variance in biological activity. This model also indicates statistical significance >99.9% with F values F = 29.321. Cross validated squared correlation coefficient of this model was 0.6481, which shows the good internal prediction power of this model. The graph of observed vs. predicted biological activities for the training and the test molecules is shown in Figure 2. In the above equations n is the number of compounds used to derive the model and values in parentheses are the 95% confidence limit of respective coefficient.The present work shows how a set of antimycobacterial activities of various 2-polyfluoroalkyl and 2-Nitrobenzylsulphanyl benzimidazoles may be treated statistically to uncover the molecular characteristics which are essential for high activity. The generated models were analyzed and validated for their statistical significance and external prediction power. A randomization test and intervariable correlation matrix were used to evaluate the possibility of “chance correlations” in the generated models. Variables in the equation revealed that thermodynamic, electronic, structural and molecular shape analysis descriptors contribute significantly to the antimycobacterial activity. The evaluation and comparison of QSAR models generated lead to the understanding that antimycobacterial growth inhibition by this diverse set of molecules correlates with the selected descriptors which could be employed for structure optimization to achieve better activity.

Acknowledgement
The authors are thankful to Vlife Science Technologies Pvt. Ltd, 1, Akshay 50, Anand Park, Aundh, Pune, India to provide trial version of software.

Reference

[1] Kazimierczuk.; Andrzejewska, M.; Kaustova, J.;Klimesova, V.; *Eur. J, Med, Chem*, **2005,** 40,203-208.
[2] W.H.O Report, Global Tuberculosis Control, Surveillance, planning, financing Geneva, **2004.**
[3] Uchel, K.H.; *Naturforch, Z.***1970,** 25b, 934-944.
[4] Wolinowska, R.; Zajdel-Dabrowska, J.; starosciak B.J.; Kazimierczuk, Z, *Acts, Microbiol, Polon,* **2002,** 51265-273.
[5] Tripathi, K. D., *Essentials of Medical Pharmacology*, 5th Edn., Jypee Brothers Medical Publishers (P) Ltd., New Delhi, **2004.**
[6] Gupta, R .A.; Gupta, A .K.; Soni, L. K.; Kaskhedikar, S. G, *Eur, J, Med ,Chem,* **2007**, 42,1109.
[7] Osman. S. A .A.; Abdalla. A .A. *J Pharm Sci*, **1983,**72, 68.
[8] Roshdi,I. M .; Ibrahim ,E. S. A, Habib, N .S, *Pharmazie*, **1976,** 31 , 856.
[9] Yardley ,V.; Snowdon ,D.; Croft, S.; Hazar. B,*Phytother Res*,**1996,** 10, 559.
[10] Bullock, F. J.; Tweedie ,J. F.; Mcritchie, D. D.; Tucker ,M .A, *J, Med, Chem*, **1970**, 13,550.
[11] Brinfworth, R. I.; Fairlie, D. P, *Biochim, Biophys Acta*,**1995,** 5,1253.
[12] Chemin, L. S.; Buisine, E.; Yardely, V.; Kohler ,S.; Debreu, M .A.; Landry, V.; Sergheraert, C.; Croft, S .L.; Karuth-seigel, L. R.; Charvet, E. D,*J , Med, Chem*, **2001**, 44, 548.
[13] Hazar ,B.; Sur, P.; Roy, D .K.; Sur ,B.; Banerjee ,A, *Planta, Med*, **1984**, 51, 295.
[14] Perry, N .B.; Blunt ,J .W,*J , Nat, Prod*, **1991**, 54, 978.
[15] V-life MDS 3.0 vlife science technologies pvt.Ltd, 1 akshay residency, plot 50, Anand park, Aundh, Pune 411007