http://www.e-journals.in
DOI:10.7598/cst2014.739

Chemical Science Transactions
2014, **3(2),** 708-716

RESEARCH ARTICLE

G-QSAR of Novel 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole Derivatives

MAYURA KALE*, GAJANAN SONWANE and RAJESH NAWALE

Department of Pharmaceutical Chemistry, Government College of Pharmacy,
Osmanpura, Aurangabad, Maharashtra, India
kale_mayura@yahoo.com

Received 6 October 2013 / Accepted 9 November 2013

Abstract: In the present work, we have applied group quantitative structure–activity relationships (G-QSAR) for exploring the relationship between the structures of a new emerging family of 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole derivatives and their antiprotozoal activities. We have developed descriptive models, in order to aid in further optimization and development of newer antiprotozoal agents containing the benzimidazole pharmacophore. G-QSAR was performed on VLife molecular design suite (MDS) 4.5 version software. The predictive power of the QSAR was checked through the cross validation technique and also by leaving some compounds as part of external test set

Keywords: GQSAR, Benzimidazole, Antiprotozoal activity, Model generation

Introduction

Chemoinformatics has been used to perform the tasks of acquiring, storing, retrieving, searching, analysing and visualizing chemical information which are required in the drug discovery process. In the initial stages of this process, extensive work is carried out to search the availability of data of molecules whose chemical structure significantly influences biological activity. The data is analysed in terms of its suitability to develop a model that can justify variation in activity in terms of physicochemical properties. After confirming the suitability of data, statistical tools are applied and relationship between activity with the descriptors is generated and termed as Quantitative Structure Activity Relationship (QSAR). The statistical methods are regression methods, which can assume linear or non-linear relationship. After validation of QSAR models, these are used for predicting biological activities of a new set of molecules that have not been synthesized up till now. These new molecules in the congeneric series can be generated in terms of virtual combinatorial library by choosing different groups at the selected sites of substitution. The generated molecules can be screened on the basis of QSAR models for deciding priority for synthesis. Also, molecules in the congeneric series can be searched on the basis of template, pharmacophore or topomer in the compound databases and then screened for deciding their procurement. Thus, QSAR is used to connect the information from molecules with known experimental activity to molecules for which newer experiments are yet to be carried out in drug discovery project[1]. In the past, the descriptors used for QSAR interrelated the chemical environment and steric properties of groups. These were considered to be independent of each other and their interactions were completely ignored. After introduction of several molecular descriptors such as topological, electro-topological and others; the current QSAR models generated using these descriptors represent properties of whole molecule rather than contributions by individual groups. These models do not clearly specify the site at which modification is required. For this purpose, 3D-QSAR models such as CoMFA were employed whose descriptors involve steric and electrostatic fields that calculated at the grid points generated around aligned set of molecules. But, as the descriptor space is very large, 3D-QSAR models are generated by using regression methods such as partial least squares (PLS) method and this can reduce the dimensionality. The 3D-QSAR models give us clues for designing new molecules by specifying areas along with its steric and electrostatic requirements of the molecules. However, one of the major drawbacks of 3D-QSAR method is its dependency on molecular alignment and conformers chosen for the alignment. This facet becomes vital when the information of bio-active conformation is absent or when molecular framework is not rigid.

From the above discussion, it is clear that there is a requirement of QSAR method, which will allow flexibility to study molecular sites of interest and capture interactions amongst them. Hence, we report herein; one such group-based QSAR (G-QSAR) method which allows ease of interpretation unlike any conventional QSAR method which could only suggest important descriptors but does not reflect the site where it has to be optimized for design of new molecules. The proposed method is tested by its application on two different data sets and their results are discussed in the following sections. In G-QSAR method, before calculating the corresponding fragment descriptors, the fragmentation of each molecule in the dataset is carried out with a set of predefined rules. In the existing methods, a predefined fragment (or group) is searched in the molecule and then it is used as a descriptor either as an indicator variable, their count or corresponding index *viz*., path count or molecular connectivity index. This method considers cross/interaction terms as descriptors to account for the fragment interactions in QSAR model while existing methods, such descriptors are not considered[2].

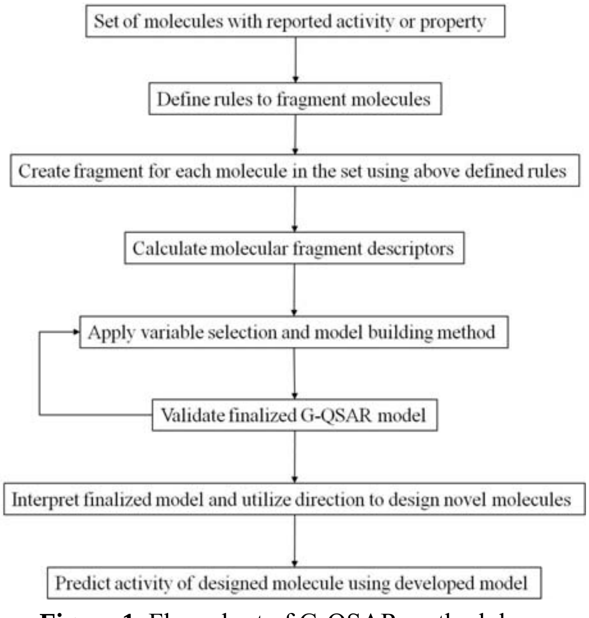

Set of molecules with reported activity or property
Define rules to fragment molecules
Create fragment for each molecule in the set using above defined rules
Calculate molecular fragment descriptors
Apply variable selection and model building method
Validate finalized G-QSAR model
Interpret finalized model and utilize direction to design novel molecules
Predict activity of designed molecule using developed model

Figure 1. Flow chart of G-QSAR methodology

Parasitic infections caused by protozoa still represent a major public health problem in developing countries. Infection by protozoa usually produces diarrhea and associated symptoms and these protozoa can also penetrate the intestinal mucosa and migrate to other organs causing severe damage. For these diseases, metronidazole and many of its analogs have been successfully used as the drugs of choice[3] for more than 40 years; however, their side effects and the development of resistant strains limit their use. Hence it is important to have more options of treatment, because of different individual responses to drugs. During the last years, an important number of benzimidazole derivatives have been synthesized and evaluated as antiprotozoals. This increasing database of benzimidazole derivatives has given access to new SAR features that can lead to the optimization of benzimidazole derivatives. The synthesis of 19 new 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole derivatives has been reported in literature. The synthesized derivatives displayed significant antiprotozoal activity high activity and selectivity. Pursuing these research consequences we have undertaken QSAR study on previously reported derivatives. The aim of the study was to identify the molecular properties which influence the antiprotozoal activity.

Experimental

A total of 20 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole derivatives have been reported to exhibit antiprotozoal activities[4]. These were used as the data set in QSAR analysis (Table 1). The IC50 (µM) values reported in the literature[5] were converted to negative logarithmic values to get pIC50 which were used for QSAR study. The molecules were divided into the training set (16 molecules) and test set (5molecules) by sphere exclusion (SE) method. The structures of the reported molecules were drawn in the 2D draw application of Molecular Design Suit (MDS) software. These 2D structures where exported to QSAR Plus window to convert 2D structures to 3D structures. After the conversion, structures were subjected to energy minimization with the help of MMFF force field and optimized molecules were used to calculate the physicochemical and alignment descriptors. In G-QSAR analysis, all the methods distributed the compounds in training set of 15 derivatives and test set of 6 derivatives. Different statistical methods like multiple linear regression (MLR)[6], partial least squares regression (PLS)[7] and Principal component regression (PCR)[8] were employed for model building.

Table 1. Chemical and biological data of 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole derivatives

Compound No.	R1	R2	R3	Actual value	pIC50 value(OA)
35	H	H	H	6.749	6.74957
36	H	Cl	H	6.768	6.768
37	H	Cl	Cl	6.852	6.852
38	CH_3	H	H	6.826	6.826
39	CH_3	H	Cl	6.892	6.8923
40	CH_3	Cl	H	6.869	6.869
41	CH_3	Cl	Cl	7.016	7.016
42	H	$COOCH_3$	H	6.955	6.955
43	H	$COOCH_3$	Cl	7.138	7.138
44	CH_3	H	$COOCH_3$	6.943	6.943
45	CH_3	Cl	$COOCH_3$	7.060	7.060
46	CH_3	$COOCH_3$	H	7.111	7.111

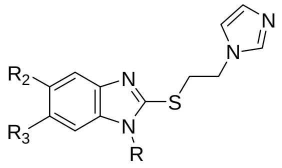
Contd...

47	CH_3	$COOCH_3$	Cl	7.033	7.033
48	H	OCH_2CH_3	H	7.144	7.14
49	H	OCH_2CH_3	Cl	7.003	7.003
50	CH_3	H	OCH_2CH_3	7.144	7.144
51	CH_3	Cl	OCH_2CH_3	7.156	7.156
52	CH_3	OCH_2CH_3	H	7.118	7.11
53	CH_3	OCH_2CH_3	Cl	7.00	7.008
Metronidazole				6.627	6.627
Albendazole				5.798	5.798

2-{[2-(1H-imidazol-1-yl)ethyl]sulfanyl}-1H-benzimidazole derivatives; R1, R2 & R3 are substituents on derivatives, OA is the observed activity

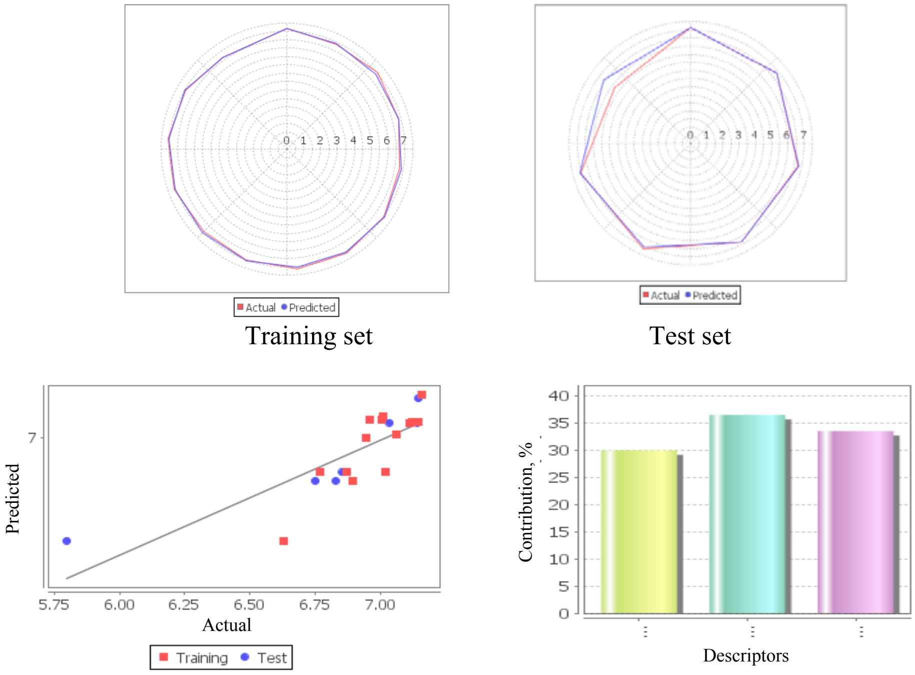
Results and Discussion

The various derivatives of 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole belong to the following parent skeleton



Interpretation of results

Model 1



Training set Test set

Fitness plot Contribution plot

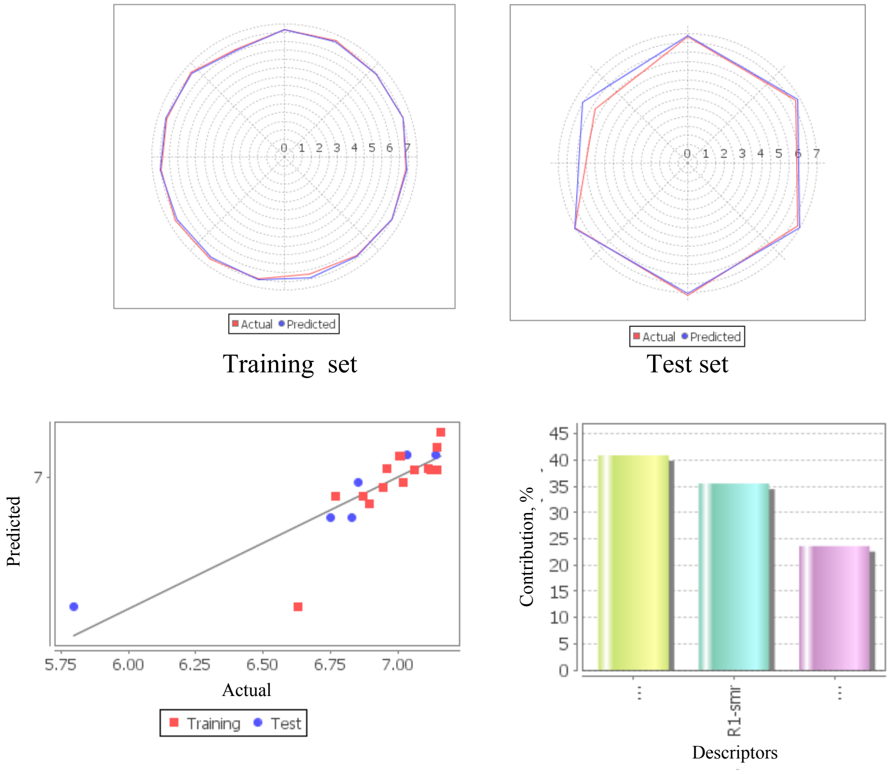
Figure 2. Model 1 (MLR method)

From the above observations, it can be seen that multiple regression (coupled with forward variable selection) has led to a statistically significant G-QSAR model. The developed G-QSAR model reveals that the descriptor[9] R1-EpsilonSS at R1 substitution position plays most important role (~38%) and is directly related in determining biological activity. The next important factor controlling changes in the activity is the log of partition coefficient at R1 substitution. This is R2-XKHydrophilicArea (~33%) which is directly proportional to the activity. Lastly, the presence of descriptor R2-SAHydrophobicArea (~30%) which is directly proportional to the activity shows the role of electronic property at R2 substitution site in determining activity.

Equation

PIc50 = 0.0050R1SAHydrophobic area+ 0.0460R1EpsilonSS+0.0055R2XRHydrophilic area+6.6555(constant). The equation explains ~75 % (r2 = 0.7587) of the total variance in the training set and has an internal (q2) and external (pred_r2) predictive ability of ~59 % and ~50%; respectively. The Fcal value of 11.5263 shows the statistical significance of 99.99% of the model which means that probability of failure of the model is 1 in 10000. In addition, the randomization test shows confidence of >99.99% (Alpha R and R2=0.00100) and that the generated model is not random and hence is chosen as the QSAR model.

Model 2



Training set

Test set

Fitness plot

Contribution plot

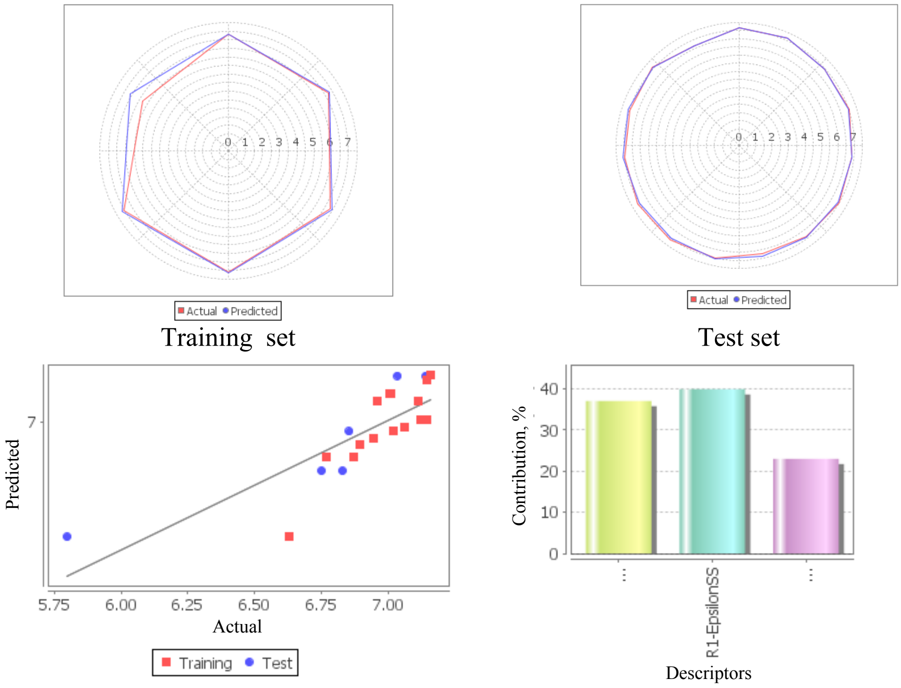
Figure 3. Model 2 (PCR method)

From the previously mentioned data, it can be seen that the principal component regression (coupled with forward variable selection) has led to a statistically significant G-QSAR model. This model reveals that the descriptor R2-SAHydrophobicArea at R1 position plays most important role (~40%) and is directly proportional in determining biological activity. Another factor affecting biological activity is the log of partition coefficient at R1 position. R1-smr (~35%) is also found to be directly related to the biological activity. Also, the presence of descriptor R2-+ve Potential Surface Area (~23%) confirms the role of electronic property at R2 substitution site in determining activity and is also directly proportional to the biological activity.

Equation

PIc50 = 0.0066R2SAHydrophobic area+ 0.0146R1smr+0.0026R2+vepotential surface area+6.5789(constant) .The equation explains ~73% (r2 = 0.7353) of the total variance in the training set and has an internal (q2) and external (pred_r2) predictive ability of ~27% and ~57%; respectively with the Fcal value 16.66. Data shows poor external validation of 27% due to which the chances of failure are more and the generated model is not random, hence is not chosen as the QSAR model.

Model 3



Training set

Test set

Fitness plot

Contribution plot

Figure 4. Model 3 (PLS method)

Partial least square (coupled with forward variable selection) method has also generated a statistically significant G-QSAR model. The developed G-QSAR model reveals that the descriptor R1-EpsilonSS at R1 substitution site plays most important role (~42%) and is directly proportional in determining biological activity. Also, log of partition coefficient at R1 position *viz*., R2-SAHydrophobicArea (~37%) is directly proportional to the activity.

Lastly, the presence of descriptor R2-XKHydrophilicArea (~23%) which is directly proportional to the activity shows the role of electronic property at R2 substitution site in determining activity. The calculated values of quantum chemical descriptors and predicted activity is shown in Table 3.

Equation

PIc50 = 0.0067R2SAHydrophobic area+ 0.0404R1EpsilonSS+0.0036R2XRHydrophilic area+6.6282(constant). The equation explains ~74 % (r2 = 0.7412) of the total variance in the training set and has an internal (q2) and external (pred_r2) predictive ability of ~62 % and ~53%; respectively. The Fcal value is 17.1803 which gives the statistical significance of 99.99% for the model which means that probability of failure of the model is 1 in 10000. In addition, the randomization test shows confidence of >99.9999% (Alpha Rand R^2=0.00100) that the generated model is not random and hence is chosen as the QSAR model. The summary of models with statistical parameters is depicted in Table 2.

Table 2. Summary of best three models developed along with statistical parameters

Method	r^2	q^2	F test	Pred r^2	Variable selection & Coefficient

MRM	0.7587	0.5958	11.526	0.5081	R2SAHydrophobicArea(0.0050(±0.002)
					R1-EpsilonSS(0.0460(±0.0109)
					R2XKHydrophilicArea(0.0055(±0.0006)
					Constant:
					6.6555

PLS	0.7412	0.6213	17.180	0.5306	R2-SAHydrophobicArea(0.0067)
					R1-EpsilonSS(0.0404)
					R2-XKHydrophilicArea(0.0036)
					Constant:
					6.6282

PCR	0.7353	0.2737	16.664	0.5798	R2-SAHydrophobicArea(0.0066) R1-smr (0.0146) R2-+vePotentialSurfaceArea**(**0.0026**)** **Constant:** 6.5789

VC is the variables counts with the value 3 obtained for all the above methods

Table 3. Calculated values of quantum chemical descriptors and predicted activity of 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole by MLR, PLS and PCR methods

Derivative	SAHydrophobic Area	EpsilonSS	XKHydrophilic Area	Predicated Value
51	48.6724	1.3	-0.121106	7.149
50	50.9416	0.6	-0.120711	7.135
44	33.1476	0.6	-0.123393	6.998
45	33.0587	1.3	-0.123302	7.013
40	28.1386	1.3	0.0943466	6.883
41	40.5796	1.3	0.266605	6.883
39	40.5796	0.6	0.266605	6.854
36	28.1386	1.3	0.0943466	6.883
53	40.5796	4.2	0.266605	7.073

Contd…

52	28.1386	4.2	0.0943466	7.055
48	28.1386	4.2	0.0943466	7.056
49	40.5796	4.2	0.266605	7.063
42	28.1386	5.2	0.0943466	7.061
46	28.1386	5.2	0.0943466	7.052
Albendazole	0	0	0	6.649
38 [b]	28.1386	0.6	0.0943466	6.854
35 [b]	28.1386	0.6	0.0943466	6.854
37 [b]	40.5796	1.3	0.266605	6.883
43 [b]	40.5796	5.2	0.266605	7.051
47 [b]	40.5796	5.2	0.266605	7.052
Mebendazole[b]	0	0	0	6.6498

SAHydrophobicArea - hydrophobic surface area, EpsilonSS - hydrogen donor capacity, XKHydrophilic Area - hydrophilic surface area, b-test set

Conclusion

A group-based quantitative structure activity relationship study was performed on a series of 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole derivatives possessing antiprotozoal activity[10] to quantify and determine those physicochemical properties that highly influence the biological activity. Two dimensional quantitative structure activity relationship (2D QSAR) study by means of multiple regression (MR) method was performed on a series of 2-{[2-(1*H*-imidazol-1-yl)ethyl]sulfanyl}-1*H*-benzimidazole derivatives possessing antiprotozoal activity using molecular design suite (VLifeMDS 4.5)[11]. This study was performed with twenty one compounds (data set) using SE algorithm, random and manual selection methods for the division of the data set into training and test sets. Multiple regression methodology with stepwise (SW) forward variable selection method was used for building the QSAR models. Statistically significant

G-QSAR models were generated. Among them most significant model has squared correlation coefficient (r^2), cross validated correlation coefficient (q^2) and predictive correlation coefficient (pred_r^2) 0.7587, 0.5958 and 0.5081; respectively The second model was generated by using PCR method. Among them, the most significant model has squared correlation coefficient (r^2), cross validated correlation coefficient (q^2) and predictive correlation coefficient (pred_r^2) 0.7353,0.2737 and 0.5798; respectively. The third model was generated by using partial least square method PLS. Among them, most significant model has squared correlation coefficient (r^2), cross validated correlation coefficient (q^2) and predictive correlation coefficient (pred_r^2)0.7412, 0.6213 and 0.5306; respectively. R1-EpsilonSS at R1 position, R2-XKHydrophilicArea and R2-SAHydrophobi Area are the three descriptors which control the biological activity and are directly related to it. In the present study, an attempt has been made to identify the necessary structural and substituent sites which can be modified so as to improve the biological activity. From the present G-QSAR analysis, three best models were generated among which any one can be used for predicting the activity of the newly designed compounds in finding some more potent molecules. Finally, it is concluded that the work presented here will play an important role in understanding the relationship of physiochemical parameters with structure and biological activity. By studying the G-QSAR model one can select the suitable substituent for further synthesizing bioactive compounds showing maximum potency.

Acknowledgement

The authors are indebted to the Principal, Government College of Pharmacy, Aurangabad for providing necessary facilities and also to Dr. Parag Choudhari for his guidance.

References

1. Ajmani S, Jadhav K and Kulkarni S, *QSAR Comb Sci.*, 2009, **28(1)**, 36-51; DOI: 10.1002/qsar.200810063
2. Gasteiger J, Chemoinformatics: A Textbook, Wiley-VCH, Weinheim, 2003.
3. Upcroft P and Upcroft J A, *Clin Microbiol Rev.*, 2001, **14(1)**, 150-164; DOI:10.1128/CMR.14.1.150-164.2001
4. Villanueva J P, Hernández A C, Lilián Y M, Méndez C C, Méndez O L, Hernández F L and Castillo R, *Bioorg Med Chem Lett.*, 2013, **23(14)**, 4221-4224; DOI:10.1016/j.bmcl.2013.05.012
5. Daisy P, Vijayalakshmi P, Selvaraj C, Singh S K and Saipriya K, Ind J Pharm Sci., 2013, **74(3)**, 217-222; DOI:10.4103/0250-474X.106063
6. Kubinyi H, QSAR: Hansch Analysis and Related Approaches, Wiley-VCH, Weinheim, 1993.
7. Geladi P and Kowalski B R, *Anal Chim Acta*, 1986, **185**, 1-17; DOI:10.1016/0003-2670(86)80028-9
8. Kleinbaum David G, Applied Regression Analysis and Multivariable Methods, Cole Publishing Company, California, 1998.
9. Wolff M, Burger’s Medicinal Chemistry and Drug Discovery, John Wiley and Sons, New York, 1995.
10. Hernández L F, Hernández C A, Castillo R, Navarrete G V, Soria O A, Hernández H M and Yépez M L, *Eur J Med Chem.*, 2010, **45(7)**, 3135-3141; DOI:10.1016/j.ejmech.2010.03.050
11. VLifeMDS 3.0, Molecular Design Suite Developed by VLife Sciences Technologies Pvt. Ltd., Pune, India, 2007.
12. Winkler D A and Burden F R, *Quant Struct Act Relat.*, 1998, **17(3)**, 224-231; DOI: 10.1002/(SICI)1521-3838(199806)17:03<224::AID-QSAR224>3.0.CO;2-6
13. Monti J M, *Life Sci.*, 1993, **53**, 1331–1338; DOI:10.1016/0024-3205(93)90592-Q
14. Weinheim, QSAR: Hansch Analysis and Related Approaches,Wiley-VCH, Weinheim, 1993.
15. Katritzky A R, Fara D, Petrukhin R, Tatham D, Maran U, Lomaka A and Karelson M, *Curr Top Med Chem.*, 2002, **2**, 1333-1356.
16. Hansch C and Lien E L, *Biochem Pharmacol.*, 1968, **17(5)**, 709-720; DOI:10.1016/0006-2952(68)90007-5
17. Bertsimas D and Tsitsiklis J, *Stat Sci.*, 1993, **8**, 10-15.