Scientific
Research
Publishing

2D-QSAR Study of a Series of Pyrazoline-Based Anti-Tubercular Agents Using Genetic Function Approximation

Hemal M. Soni[1], Popatbhai K. Patel[2], Mahesh T. Chhabria[3], Dharmraj N. Rana[4], Bhushan M. Mahajan[1], Pathik S. Brahmkshatriya[4]*

[1]M/S Piramal Enterprises Ltd., Piramal Enterprises Limited-Discovery Solutions, Plot No. 18, Pharmaceutical Special Econonic Zone, Village Matoda, Ta. Sanand, Ahmedabad, India
[2]M.G. Science Institute, Dada Saheb Mavlankar Campus, Opp. Gujarat University, Ahmedabad, India
[3]Department of Pharmaceutical Chemistry, L. M. College of Pharmacy, Ahmedabad, India
[4]Oxygen Healthcare Res. Pvt. Ltd., Plot No. 35, Panchratna Industrial Estate, Near IBP Laxminarayan Petrol Pump, Changodar, Sarkhej-Bavla Road, Ahmedabad, India
Email: *pathik26@yahoo.com

Abstract

A series of pyrazoline-based new heterocycles have recently been synthesized from our group where some of the compounds display potent anti-tubercular activity against *Mycobacterium tuberculosis* H37Rv. In order to further explore the potency of the compounds, quantitative structure activity relationship study is carried out using genetic function approximation. Statistically significant ($r^2 = 0.85$) and predictive $\left(r^2_{\text{pred}} = 0.89 \text{ and } r^2_m = 0.74 \right)$ QSAR models are developed. It is evident from the QSAR study that majority of the anti-tubercular activity is found to be driven by lipophilicity. Also, molecular solubility, Jurs and shadow descriptors influence the biological activity significantly. Also, positive contribution of molecular shadow descriptors suggests that molecules with bulkier substituents are more likely to enhance anti-tubercular activity. Since the developed QSAR models are found to be statistically significant and predictive, they potentially can be applied for predicting anti-tubercular activity of new molecules for prioritization of molecules for synthesis.

Keywords

QSAR, Genetic Function Approximation, Descriptors, Cross-Validation

---

*Corresponding author.

1. Introduction

World Health Organization (WHO) estimates that almost one-third of the world's population, (~2 billion people) is infected with the tuberculosis [1]. Every year, more than 8 million people develop an active form of this disease, which claims the lives of nearly 2 million. WHO estimated in 2002 that if the worldwide spread of tuberculosis was left unchecked, it would be responsible for nearly 36 million more deaths by 2020. Effective and specific anti-tubercular drugs are still not found and classical antibiotics are currently being used for curing tuberculosis. However, effectiveness of such treatment is rather controversial [2]. Multidrug-resistant TB (MDR-TB), a form of TB that does not respond to the first-line TB drugs and extensively drug-resistant TB (XDR-TB), an MDR-TB with resistance to aminoglycosides and fluoroquinolones has become a serious threat to control and treatment of tuberculosis. There are also a few cases reported of totally drug resistant tuberculosis (TDR-TB); which has raised alarming concerns on the existing drug regimen [3]. This implies urgent need to discover newer anti-tubercular agents with newer molecular mechanisms.

Quantitative structure activity relationship (QSAR) is one of the most widely used tools to design newer candidates for several therapeutic areas [4]-[6]. It provides useful insights into the structural features which are responsible for the biological activity and help to generate a mathematical model which can predict activity of untested compounds quantitatively. QSAR study usually leads to a predictive formula by correlation of physicochemical properties of a congeneric series with the biological activity [7].

Earlier from our laboratory, a series of pyrazoline-based benzoxazoles are identified as potent anti-tubercular agents [8] [9]. In order to further investigate the potency of the molecule as a part of lead optimization program, we carry out QSAR study by using Genetic Function Approximation (GFA) technique [10]. GFA algorithm is a novel approach to create structure—activity models. It searches QSAR models automatically by combining statistical modeling with genetic algorithm tools. Typically, thousands of candidate models are generated and tested during evolution. However, only the superior (best) models survive; which are used as "parents" to create the next generation of candidate models. Previously, we have successfully applied GFA to generate a variety of QSAR models [4] [5]. Such models provide useful structure-activity insights, which can be used for prioritization of synthetic efforts to generate and lead optimization strategies.

2. Experimental

2.1. Data Set

In present studies, a series of substituted pyrazoline-based compounds reported by Rana *et al.* as potent anti-tubercular agents was selected [8] [9]. Fifty four compounds were randomly divided into training and test set, the former set consisting of thirty nine compounds and the remaining fifteen compounds were taken as the test set. Structures of all the compounds used for 2D-QSAR analysis and their anti-tubercular activity (MIC, µg/mL) are given in **Table 1**. For all the compounds, the experimental values of biological activity (MIC) are used in the negative logarithmic scale (pMIC) to achieve normal distribution. Structures of all compounds were sketched by using visualizer module of Discovery Studio 2.1 software (Accelrys Inc., USA). CHARMM force field was used for the calculation of potential energy. Energy minimization of all the compounds was done using Smart Minimizer method until the root mean square (RMS) gradient value becomes smaller than 0.001 kcal/mol Å. This was followed by geometry optimization by semi empirical MOPAC-AM1 method (Astin Method-1).

2.2. Descriptor Calculation

"Calculate Molecular Properties" protocol of the Discovery Studio 2.1 was used to calculate various physicochemical descriptors like structural, thermodynamic, steric, electronic and quantum mechanical descriptors. Further, a correlation matrix of the molecular descriptors was generated and highly correlated descriptors with a correlation value of 0.6 or above were discarded from the study. Remaining least correlated descriptors were used to develop 2D-QSAR models. Descriptors included in developing 2D-QSAR models are listed and described in **Table 2**.

2.3. Regression Analysis

The advantage of GFA is that the data set is being modeled to generate a population of equations rather than one

Table 1. Chemical structures and biological activity of training set (**1-39**) and test set (**40-54**) compounds.

Comp. No.	X	R_1	R_2	R_3	R_4	R_5	R_6	pMIC
1	-	OCH$_3$	H	H	H	H	H	7.114
2	-	OCH$_3$	Cl	H	Cl	H	H	8.801
3	-	OCH$_3$	H	H	F	H	H	7.438
4	-	OCH$_3$	Br	H	H	H	H	7.508
5	-	H	H	H	H	H	H	7.072
6	-	H	Br	H	H	H	H	7.475
7	-	Cl	H	H	H	H	H	8.007
8	-	Cl	Br	H	H	H	H	7.514
9	NH$_2$	OCH$_3$	H	H	H	H	H	7.447
10	NH$_2$	OCH$_3$	Cl	H	Cl	H	H	7.526
11	NH$_2$	OCH$_3$	H	H	F	H	H	7.770
12	NH$_2$	OCH$_3$	Br	H	H	H	H	7.535
13	NH$_2$	H	H	H	H	H	H	7.107
14	NH$_2$	H	Br	H	H	H	H	7.202
15	NH$_2$	Cl	H	H	H	H	H	7.152
16	NH$_2$	Cl	Br	H	H	H	H	8.125
17	SH	OCH$_3$	H	H	H	H	H	7.769
18	SH	OCH$_3$	Cl	H	Cl	H	H	7.845
19	SH	OCH$_3$	H	H	F	H	H	7.489
20	SH	OCH$_3$	Br	H	H	H	H	8.852
21	SH	H	H	H	H	H	H	7.130
22	SH	H	Br	H	H	H	H	7.521
23	SH	Cl	H	H	H	H	H	8.474
24	SH	Cl	Br	H	H	H	H	7.857
25	SH	H	H	OCH$_3$	OCH$_3$	H	H	7.502
26	SH	H	H	H	OCH$_3$	H	H	7.769
27	SH	H	Cl	H	H	H	H	7.775
28	SH	H	H	Cl	H	H	H	8.059
29	SH	H	H	H	Cl	H	H	8.059
30	SH	OCH$_3$	H	OCH$_3$	OCH$_3$	H	H	8.835
31	SH	OCH$_3$	H	H	OCH$_3$	H	H	7.502
32	SH	OCH$_3$	Cl	H	H	H	H	7.808
33	SH	OCH$_3$	H	Cl	H	H	H	7.206
34	SH	OCH$_3$	H	H	Cl	H	H	8.092
35	SH	OCH$_3$	OCH$_3$	H	H	H	OCH$_3$	8.835
36	SH	OCH3	H	Br	H	H	H	8.551
37	SH	OCH3	H	H	Br	H	H	8.136
38	NH$_2$	H	H	OCH3	OCH3	H	H	7.784

Continued

39	NH$_2$	H	H	H	OCH$_3$	H	H	7.447
40	NH$_2$	H	Cl	H	H	H	H	7.152
41	NH$_2$	H	H	Cl	H	H	H	7.152
42	NH$_2$	H	H	H	Cl	H	H	7.754
43	NH$_2$	OCH$_3$	Cl	H	H	H	H	7.489
44	NH$_2$	OCH$_3$	H	H	Br	H	H	7.234
45	NH$_2$	H	H	OCH$_3$	OCH$_3$	H	H	8.135
46	NH$_2$	H	H	H	OCH$_3$	H	H	7.519
47	NH$_2$	H	Cl	H	Cl	H	H	7.258
48	NH$_2$	H	H	H	Cl	H	H	7.524
49	NH$_2$	H	H	H	F	H	H	7.205
50	SH	H	H	OCH$_3$	OCH$_3$	H	H	8.151
51	SH	H	H	H	OCH$_3$	H	H	7.838
52	SH	H	Cl	H	Cl	H	H	7.575
53	SH	H	H	H	Cl	H	H	7.843
54	SH	H	H	H	F	H	H	7.825

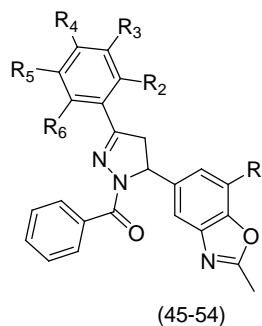

(1-8)

(9-24)

(25-44)

(45-54)
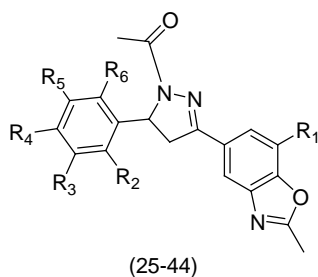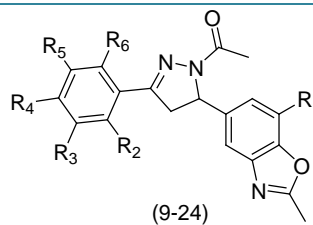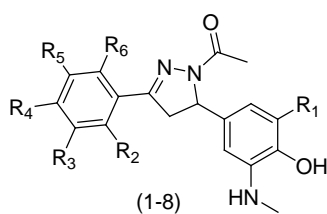
Table 2. List of descriptors used in the study.

Sr. No.	Descriptor	Definition
1	ALogP	Log of the octanol-water partition coefficient using Ghose and Crippen's method
2	Jurs_RNCG	Charge of most negative atom divided by the total negative charge
3	Apol	The sum of the atomic polarizabilities
4	Jurs_DPSA_1	Partial positive solvent-accessible surface area minus partial negative solvent-accessible surface area
5	Shadow_XZ	Area of the molecular shadow in the xz plane
6	Molecular_Solubility	Molecular solubility expressed as LogS, where S is the solubility in mol/L

single equation for descriptor-activity correlation. GFA is genetic principle based method of variable selection, which combines Holland's genetic algorithm and Friedman's multivariate adaptive regression splines. Thus, it evolves the population of equations that best fit the training set data.

In GFA, a particular number of equations (set at 100 by default) are randomly generated. The pairs of "parent" equations then are chosen randomly from this set of 100 equations. After this, "crossover" operations are performed at random. The number of crossing over was set at 5000 by default. The goodness of each progeny equation is assessed by Friedman's lack of fit (LOF) score

$$\text{LOF} = \text{LSE} \Big/ \big[1 - (c + dp)/m\big]^2$$

where c is the number of basis functions in the model, LSE is the least-squares error, p is the number of descriptors, d is smoothing parameter, and m is the number of observations in the training set. The smoothing parameter controls the scoring bias between equations of different sizes. It was set at default value of 0.5. GFA crossover of 5000 was set to give reasonable convergence. The length of equation was fixed to six terms, the population size was established as 100, and the mutation probability was specified as 0.1. Best three equations, out of the 100 equations, were chosen based on the statistical parameters like LOF, regression coefficient (r), adjusted regression coefficient (r_{adj}), cross-validated regression coefficient (r_{cv}) and F-test values.

2.4. Validation Test

Variance inflation factor (VIF) analysis was performed to check the inter-correlation of descriptors. VIF value is calculated from $1/1 - r^2$, where r^2 is the multiple correlation coefficient of one molecular descriptor's effect regressed on the remaining descriptors. VIF value greater than 10 suggests chance-correlation and hide the information of molecular descriptors by inter-correlation of descriptors [11].

It is proven that a high value of statistical characteristics r and F and low value of s and LOF need not be the criteria of a highly predictive model. Thus, in order to evaluate the predictive ability of the 2D-QSAR model, the external predictability method described by Roy *et al.* was used [12]. It was determined by calculating the value of predictive r^2 $\left(r_{\text{pred}}^2\right)$ using the following equation

$$r_{\text{pred}}^2 = 1 - \frac{\sum \left(Y_{\text{Pred(test)}} - Y_{\text{Obs(test)}}\right)^2}{\sum \left(Y_{\text{Obs(test)}} - \overline{Y}_{\text{training}}\right)^2}$$

where, $Y_{\text{Obs(test)}}$ and $Y_{\text{Pred(test)}}$ are the observed and predicted activity values, respectively, of the test set compounds and Y_{training} is the mean activity value of the training set.

3. Results and Discussion

In the present study, 31 descriptors were selected initially for correlation with anti-tubercular activity. The 31 preselected descriptors represented different class of descriptors such as quantum mechanical, steric, geometric, thermodynamic, and electronic. The descriptors were correlated with training set using GFA methodology. Initially, 100 2D-QSAR equations with six descriptors were generated. The results of the best three models are given in **Table 3** along with their regression statistics.

For a statistically significant model, it is inevitable that the descriptors evolved in the equation should be least inter-correlated with each other. In the present study, the inter-correlation of the descriptors used in the selected models was found to be very low. The correlation matrix for the used descriptors is shown in **Table 4**.

Further to check the inter-correlation of descriptors, variance inflation factor (VIF) analysis was performed (as described in Section 2.4). VIF values of these descriptors were found to be 2.010 (ALogP), 1.243 (Jurs_RNCG), 2.558 (Apol), 1.366 (Jurs_DPSA_1), 1.520 (Shadow_XZ) and 1.585(Molecular_Solubility). All the VIF values were found to be less than 10, which suggest very less multi-collinearity within descriptors. The models were also evaluated for their predictive power, *i.e.* internal and external cross-validation. The results for Equation (1) are summarized in **Table 5** and **Table 6**.

Figure 1 and **Figure 2** show the plot of observed Vs predicted activity for training and test set compounds, respectively as per Equation (1). It was seen that the models displayed r_{pred}^2 and r_m^2 in the acceptable range [12].

The descriptors used in the study were found to have significant influence on the biological activity as seen

Table 3. Selected 2D-QSAR equations and their regression statistics.

Eq. No.	Description	r^2	r^2_{adj}	r^2_{cv}	r^2_{pred}	r^2_m	LOF	F
1	pMIC = 18.353 + 16.157 * AlogP − 1.179 * Jurs_RNCG + 0.062 * Apol − 1.366 * Jurs_DPSA_1 + 0.685 * Shadow_XZ + 3.233 (Molecular_Solubility)	0.851	0.739	0.724	0.894	0.742	0.144	13.647
2	pMIC = 18.410 + 11.157 * AlogP − 0.989 * Jurs_RNCG + 0.098 * Apol − 1.354 * Jurs_DPSA_1 + 4.548 * Shadow_XZ + 2.589 (Molecular_Solubility)	0.811	0.701	0.718	0.865	0.713	0.156	13.297
3	pMIC = 18.473 + 16.589 * AlogP − 1.185 * Jurs_RNCG + 0.121 * Apol − 1.987 * Jurs_DPSA_1 + 2.668 * Shadow_XZ + 2.936 (Molecular_Solubility)	0.794	0.678	0.699	0.814	0.667	0.204	12.842

**Table 4.** Correlation matrix of the descriptors used in the equations.

	pMIC	AlogP	Jurs_RNCG	Apol	Jurs_DPSA_1	Shadow_XZ	Molecular_Solubility
pMIC	1						
AlogP	0.192	1					
Jurs_RNCG	0.317	0.582	1				
Apol	0.211	0.132	0.449	1			
Jurs_DPSA_1	0.336	0.304	0.480	0.197	1		
Shadow_XZ	0.168	0.256	0.281	0.119	0.297	1	
Molecular_Solubility	0.089	0.244	0.378	0.090	0.269	0.014	1
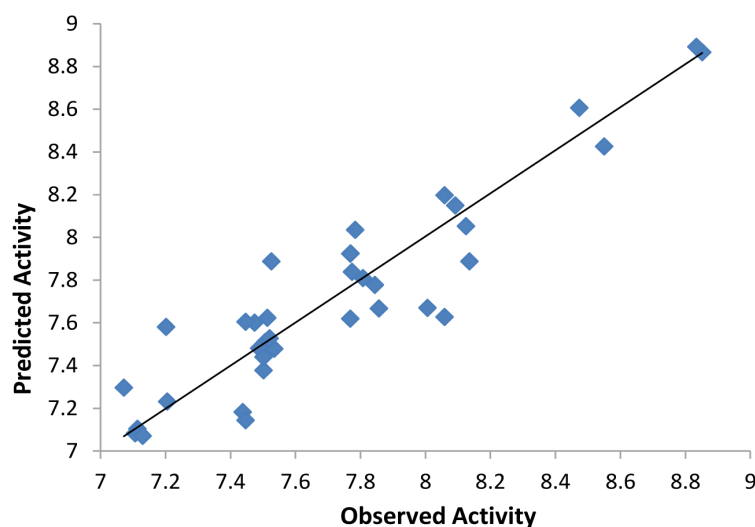
Figure 1. Plot of observed Vs predicted pMIC values of training set compounds (as per Equation (1)).

from their high coefficients values. Noticeably, the activity was found to be governed chiefly through lipophilicity (AlogP). As seen from the positive coefficient, lipophilicity positively influenced the activity. Indeed, compounds with halogens (bromo/chloro, **2**, **7**, **16**, **20**) were found to possess high anti-tubercular activity whereas compounds with polar groups (**9-15**) were found to be less active. Jurs descriptors are a group of molecular descriptors which combine electronic and shape information to characterize molecules [13]. They are calculated by mapping atomic partial charges on solvent-accessible surface areas of individual atoms. Jurs_RNCG is charge of most negative atom divided by the total negative charge. Jurs_DPSA_1 is partial positive solvent-accessible surface area minus partial negative solvent-accessible surface area. A critical analysis of the generated equations

Table 5. Observed and predicted pMIC values of training set compounds (as per Equation (1)).

Compound No.	pMIC (observed)	pMIC (predicted)	Residual
1	7.114	7.103	−0.011
2	8.801	9.049	0.248
3	7.438	7.182	−0.256
4	7.508	7.509	0.001
5	7.072	7.296	0.224
6	7.475	7.600	0.125
7	8.007	7.669	−0.338
8	7.514	7.622	0.108
9	7.447	7.143	−0.304
10	7.526	7.887	0.361
11	7.770	7.924	0.154
12	7.535	7.478	−0.057
13	7.107	7.082	−0.025
14	7.202	7.580	0.378
15	7.152	6.957	−0.195
16	8.125	8.052	−0.073
17	7.769	7.923	0.154
18	7.845	7.776	−0.069
19	7.489	7.480	−0.009
20	8.852	8.865	0.013
21	7.130	7.070	−0.060
22	7.521	7.527	0.006
23	8.474	8.605	0.131
24	7.857	7.666	−0.191
25	7.502	7.377	−0.125
26	7.769	7.618	−0.151
27	7.775	7.838	0.063
28	8.059	7.626	−0.433
29	8.059	8.196	0.137
30	8.835	8.891	0.056
31	7.502	7.439	−0.063
32	7.808	7.809	0.001
33	7.206	7.231	0.025
34	8.092	8.148	0.056
35	8.835	9.019	0.184
36	8.551	8.425	−0.126
37	8.136	7.887	−0.249
38	7.784	8.034	0.250
39	7.447	7.604	0.157
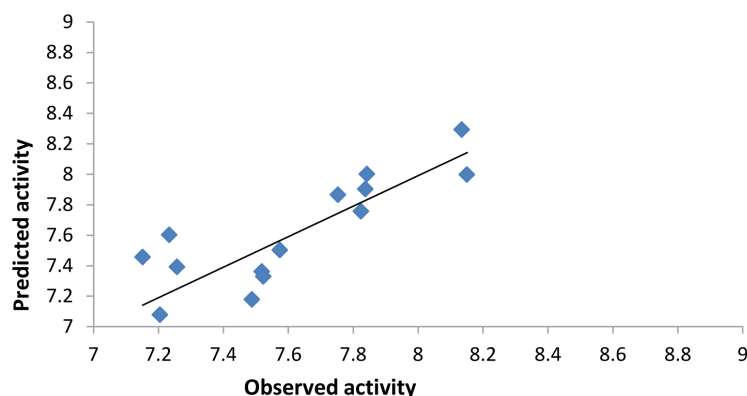
Figure 2. Plot of observed Vs predicted pMIC values of test set compounds (as per Equation (1)).

Table 6. Observed and predicted pMIC values of test set compounds (as per Equation (1)).

Compound No.	pMIC (observed)	pMIC (predicted)	Residual
40	7.152	7.456	0.304
41	7.152	6.808	−0.344
42	7.754	7.865	0.111
43	7.489	7.178	−0.311
44	7.234	7.602	0.368
45	8.135	8.293	0.158
46	7.519	7.361	−0.158
47	7.258	7.392	0.134
48	7.524	7.329	−0.195
49	7.205	7.078	−0.127
50	8.151	7.997	−0.154
51	7.838	7.902	0.064
52	7.575	7.502	−0.073
53	7.843	8.000	0.157
54	7.825	7.756	−0.069

suggested negative contribution of these descriptors on biological activity. This means that the charge distribution within the molecules serves as the driving force for intermolecular interactions and the higher the relative charge the smaller the interactions. The above fact is exemplified from compounds **2**, **20**, **30** where lower values of the above descriptors resulted in increase in activity. Another set of geometrical descriptors, Molecular Shadow descriptors like Shadow_XZ (area of the molecular shadow in the XZ plane) also showed significant contribution to anti-tubercular activity with the coefficient being positive. This shows that molecules with bulkier substituents (**2**, **20**, **30**, **35**) are more likely to show activity. In consistent with the above correlation, compounds **1**, **5**, **13** and **21** (with one or more H substituents) stood out as less active due to low values of Shadow_XZ. Apol (the sum of the atomic polarizabilities) also contributed positively to anti-tubercular activity. However, its low co-efficient signals its low contribution as compared to the other descriptors.

4. Conclusion

Developed 2D-QSAR models were found to be statistically significant as seen from their regression statistics.

Also, during internal and external cross-validation studies, very low residuals were obtained which suggested that developed models were predictive. This was also supported by their satisfactory r_{pred}^2 and r_m^2 values. Anti-tubercular activity of this series of compounds was found to be governed chiefly by lipophilicity. Compounds with polar substitution were found to be less active. Shadow and Jurs descriptors also positively influenced the activity; compounds with bulkier substitutions were generally found to be potent. These results would provide valuable guidance for improving the anti-tubercular activity of pyrazoline-based compounds.

References

[1] Dye, C., Lönnroth, K., Jaramillo, E., Williams, B.G. and Raviglione, M. (2009) Trends in Tuberculosis Incidence and Their Determinants in 134 Countries. *Bull*, **87**, 683-691. http://dx.doi.org/10.2471/BLT.08.058453

[2] Cole, S.T. and Riccardi, G. (2011) New Tuberculosis Drugs on the Horizon. *Current Opinion in Microbiology*, **14**, 570-576. http://dx.doi.org/10.1016/j.mib.2011.07.022

[3] Udwadia, Z. F., Amale, R.A., Ajbani, K.K. and Rodrigues, C. (2012) Totally Drug-Resistant Tuberculosis in India. *Clinical Infectious Diseases*, **54**, 579-581. http://dx.doi.org/10.1093/cid/cir889

[4] Chhabria, M.T., Mahajan, B.M. and Brahmkshatriya, P.S. (2011) QSAR Study of a Series of Acyl Coenzyme A (CoA): Cholesterol Acyltransferase Inhibitors Using Genetic Function Approximation. *Medicinal Chemistry Research*, **20**, 1573-1580. http://dx.doi.org/10.1007/s00044-010-9413-3

[5] Buha, V.M., Rana, D.N., Chhabria, M.T., Chikhalia, K.H., Mahajan, B.M., Brahmkshatriya, P.S. and Shah, N.K. (2013) Synthesis, Biological Evaluation and QSAR Study of a Series of Substituted Quinazolines as Antimicrobial Agents. *Medicinal Chemistry Research*, **22**, 4096-4109. http://dx.doi.org/10.1007/s00044-012-0408-0

[6] Tropsha, A. (2010) Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, **29**, 476-488. http://dx.doi.org/10.1002/minf.201000061

[7] Ertan, T., Yildiz, I., Tekiner-Gulbas, B., Bolelli, K., Temiz-Arpaci, O., Ozkan, S., Yalcin, I. and Aki, E. (2009) Synthesis, Biological Evaluation and 2D-QSAR Analysis of Benzoxazoles as Antimicrobial Agents. *European Journal of Medicinal Chemistry*, **44**, 501-510. http://dx.doi.org/10.1016/j.ejmech.2008.04.001

[8] Rana, D.N., Chhabria, M.T., Shah, N.K. and Brahmkshatriya, P.S. (2014) Pharmacophore Combination as a Useful Strategy to Discover New Antitubercular Agents. *Medicinal Chemistry Research*, **23**, 370-381. http://dx.doi.org/10.1007/s00044-013-0645-x

[9] Rana, D.N., Chhabria, M.T., Shah, N.K. and Brahmkshatriya, P.S. (2014) Discovery of New Antitubercular Agents by Combining Pyrazoline and Benzoxazole Pharmacophores: Design, Synthesis and Insights into the Binding Interactions. *Medicinal Chemistry Research*, **23**, 2218-2228. http://dx.doi.org/10.1007/s00044-013-0815-x

[10] Rogers, D. and Hopfinger, A.J. (1994) Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *Journal of Chemical Information and Computer Sciences*, **34**, 854-866. http://dx.doi.org/10.1021/ci00020a020

[11] Jaiswal, M., Khadikar, P.V., Scozzafava, A. and Supuran, C.T. (2004) Carbonic Anhydrase Inhibitors: The First QSAR Study on Inhibition of Tumor-Associated Isoenzyme IX with Aromatic and Heterocyclic Sulfonamides. *Bioorganic & Medicinal Chemistry Letters*, **14**, 3283-3290. http://dx.doi.org/10.1016/j.bmcl.2004.03.099

[12] Roy, P.P. and Roy, K. (2007) On Some Aspects of Variable Selection for Partial Least Squares Regression Models. *QSAR & Combinatorial Science*, **27**, 302-313.

[13] Stanton, D.T. and Jurs, P.C. (1990) Development and Use of Charged Partial Surface Area Structural Descriptors in Computer Assissted Quantitative Structure Property Relationship Studies. *Analytical Chemistry*, **62**, 2323-2329. http://dx.doi.org/10.1021/ac00220a013