International Journal of ChemTech Research
CODEN (USA): IJCRGG ISSN : 0974-4290
Vol.4, No.1, pp 41-47, Jan-Mar 2012

QSAR Modeling of Antitmycobacterial activities of N-Benzylsalicylamides and N-Benzylsalicylthioamides derivatives against *Mycobacterium kansasii* CNCTC My (6509/96) using stepwise and PLS Method

Supratim Ray

Division of Pharmaceutical Chemistry, Dr. B C Roy College of Pharmacy & Allied Health Sciences, Bidhannagar, Durgapur, 713 206, India.

Corres. Author: supratimray_in@yahoo.co.in
Ph: +91343-2532678, Telefax: +91343-2532679

Abstract: This study gives a quantitative structure activity relationship (QSAR) correlation of the fourty four N-Benzylsalicylamides and N-Benzylsalicylthioamides derivatives properties reported by Dolezal et al against *Mycobacterium kansasii* CNCTC My (6509/96). The reported minimum inhibitory concentrations [MIC] of the compounds determined after 14 days of incubation. The study was performed using electrotopological state atom (E-state) parameter as descriptors. Different statistical tools used in this communication are stepwise regression analysis and partial least squares analysis (PLS). Based on internal validation (Q^2) stepwise regression analysis (Q^2= 0.6224) and on the basis of external validation (R^2_{pred}) PLS analysis was found to be the best model (R^2_{pred}=0.8057).
Key words: QSAR, E-state, stepwise regression, PLS, N-Benzylsalicylamides, Benzylsalicylthioamides.

Introduction

The compounds having property of inhibiting the growth of mycobacteria is important due to its role in human infection. The most common disease produced by of *Mycobacterium kansasii* infection is a chronic pulmonary infection that resembles pulmonary tuberculosis. However, it may also infect other organs. *M kansasii* infection is the second-most-common nontuberculous opportunistic mycobacterial infection associated with AIDS (1). Along with these the emergence of antibiotic-resistant pathogen agents is a serious health problem worldwide today. Due to emergence of multidrug resistance of the drugs, there is an urgent need for the development of new drug candidate as well as gaining further (and deeper) knowledge of the mechanisms of action of existing (and future) active compounds. In this context a QSAR study was performed to the antimycobacterial activities of two moieties N-Benzylsalicylamides and N-Benzylsalicylthioamides derivatives against *Mycobacterium kansasii* CNCTC My (6509/96).

Materials and Methods

The Data-set and descriptors

The *in vitro* antimycobacterial activities of N-Benzylsalicylamides and N-Benzylsalicylthioamides derivatives against *Mycobacterium kansasii* CNCTC My (6509/96) were reported by Dolezal et al (2) were

used as the model data-set for the present QSAR analysis (Table 1). The reported minimum inhibitory concentrations [MIC] of the compounds determined after 14 days of incubation were in μM range which was converted to mM range and then to logarithmic scale [log (10^3 / MIC)]. The QSAR analysis was performed using electrotopological state atom (E-state) parameter. The whole data set contain fourty four compounds and all the compounds contain 17 common atoms (excluding hydrogen). The atoms of the molecules were numbered keeping serial numbers of the common atoms same in all the compounds (as shown in Fig. 1). The electrotopological states of the 17 common atoms for all of the compounds were found out using a VISUAL BASIC program SRETSA developed partly by the author (3). The program uses, as input, only the connection table in a specific format along with intrinsic state values of different atoms. To the output file thus obtained, the biological activity data were introduced to make it ready for subsequent regression analysis.

Model development

To begin the model development process, the whole data set (n=44) was divided into training (n=33, 75% of the total number of compounds) and test (n=11, 25% of the total number of compounds) sets by *k*-means clustering technique (4) applied on standardized descriptor matrix of the E-state parameters. QSAR models were developed using the training set compounds (optimized by Q^2), and then the developed models were validated (externally) using the test set compounds. The stepwise regression and PLS were performed using statistical software MINITAB (5).

Stepwise Regression

In stepwise regression (6), a multiple term linear equation was built step-by-step. The basic procedures involve (1) identifying an initial model, (2) iteratively "stepping", i.e., repeatedly altering the model of the previous step by adding or removing a predictor variable in accordance with the "stepping criteria", (F = 4 for inclusion; F = 3.9 for exclusion) in our case and (3) terminating the search when stepping is no longer possible given the stepping criteria, or when a specified maximum number steps has been reached. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the equation. That variable will then be included in the model, and the process started again. A limitation of the stepwise regression search approach is that it presumes that there is a single "best" subset of X variables and seeks to identify it. There is often no unique "best" subset, and all possible regression models with a similar number of X variables as in the stepwise regression solution should be fitted subsequently to study whether some other subsets of X variables might be better.

PLS

PLS is a generalization of regression, which can handle data with strongly correlated and/or noisy or numerous X variables (7, 8). It gives a reduced solution, which is statistically more robust than MLR. The linear PLS model finds "new variables" (latent variables or X scores) which are linear combinations of the original variables. To avoid over fitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are nonsignificant. Application of PLS thus allows the construction of larger QSAR equations while still avoiding over fitting and eliminating most variables. PLS is normally used in combination with cross validation to obtain the optimum number of components. This ensures that the QSAR equations are selected based on their ability to predict the data rather than to fit the data. In case of PLS analysis on the present data set, based on the standardized regression coefficients, the variables with smaller coefficients were removed from the PLS regression until there was no further improvement in Q^2 value irrespective of the components.

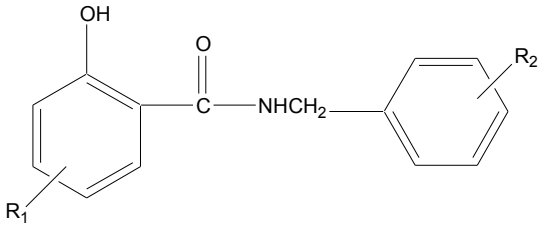
Statistical qualities

The statistical qualities of the equations were judged by the parameters such as *determination coefficient* (R^2) and *variance ratio* (F) at specified *degrees of freedom* (*df*) (9). The generated QSAR equations were validated by leave-one-out *cross-validation* R^2 (Q^2) and *predicted residual sum of squares* (*PRESS*) (10, 11) and then were used for the prediction of antimycobacterial activity of the test set compounds. The prediction qualities of the models were judged by statistical parameters like predictive R^2 (R^2_{pred}).

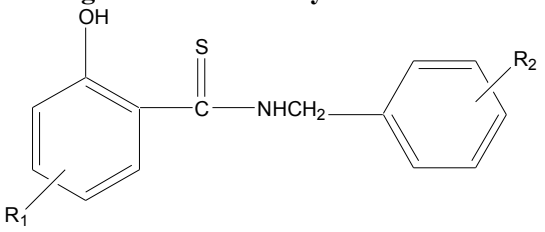
Results and Discussion

Membership of compounds in different clusters generated using *k*-means clustering technique is shown in Table 2. The test set size was set to approximately 25% to the total data set size (12) and the test set members along with their observed and calculated activity are given in Table 3. Statistical qualities of all important models are listed in Table 4. The results obtained from different statistical methods are described below and the interpretations of the equations are also depicted.

Table 1: Molecular scaffolds of the compounds along with their activity



N-Benzylsalicylamides (I)

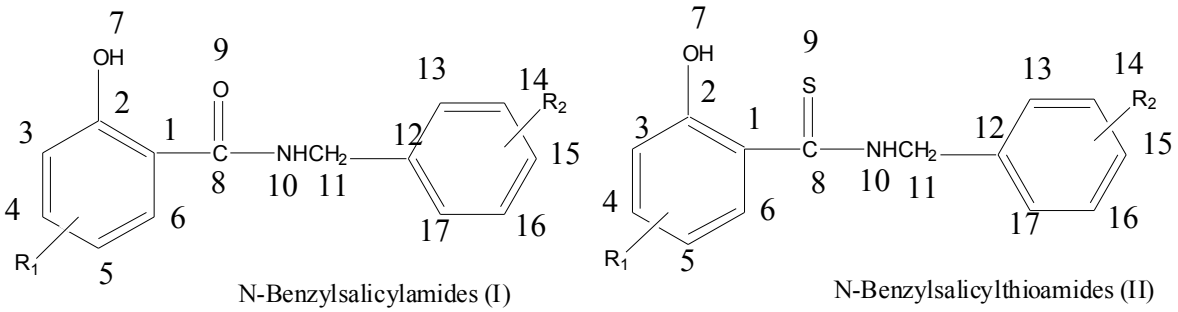


N-Benzylsalicylthioamides (II)

Compound No.	Type of compound	R_1	R_2	MIC value (µmol/L) against *Mycobacterium kansasii* CNCTC My (6509/96) after 14 days (C_{14d})	pC_{14d} = Log ($1000/C_{14d}$)
1	I	H	4-tert-but	32	1.49485
2	I	H	3-CF_3	62.5	1.20412
3	I	5-Br	3-Br	32	1.49485
4	I	5-Br	4-Br	32	1.49485
5	I	3,5 Cl_2	4-tert-but	62.5	1.20412
6	I	4-Cl	4-Br	32	1.49485
7	I	4-CH_3	H	125	0.90309
8	I	4-CH_3	4-CH_3	250	0.60206
9	I	4-CH_3	4-Cl	250	0.60206
10	I	4-CH_3	4-tert-but	62.5	1.20412
11	I	4-CH_3	3-NO_2	62.5	1.20412
12	I	4-OCH_3	3-Cl	62.5	1.20412
13	I	3-CH_3	H	62.5	1.20412
14	I	3-CH_3	4-Cl	62.5	1.20412
15	I	3,5 Br_2	4-CF_3	62.5	1.20412
16	II	H	H	2	2.69897
17	II	H	4-CH_3	0.5	3.30103
18	II	H	4-Cl	0.5	3.30103
19	II	H	4-OCH_3	4	2.39794
20	II	H	3,4 Cl_2	2	2.69897
21	II	H	4-F	2	2.69897
22	II	H	3-CH_3	2	2.69897
23	II	H	4-tert-but	2	2.69897
24	II	H	3-Cl	1	3
25	II	H	3-CF_3	2	2.69897
26	II	5-Br	3,4 Cl_2	16	1.79588
27	II	5-Br	3-Br	8	2.09691
28	II	5-Br	4-Br	4	2.39794
29	II	5-Cl	H	8	2.09691
30	II	5-Cl	3,4 Cl_2	16	1.79588
31	II	5-Cl	4-F	8	2.09691
32	II	3,5 Cl_2	3,4 Cl_2	32	1.49485
33	II	3,5 Cl_2	4-tert-but	62.5	1.20412
34	II	4-Cl	4-Br	4	2.39794
35	II	4-CH_3	H	1	3
36	II	4-CH_3	4-CH_3	0.5	3.30103
37	II	4-CH_3	4-Cl	1	3
38	II	4-CH_3	4-tert-but	1	3
39	II	4-CH_3	3-NO_2	4	2.39794
40	II	5-OCH_3	H	16	1.79588

41	II	4-OCH_3	H	8	2.09691
42	II	4-OCH_3	3-Cl	1	3
43	II	3-CH_3	4-Cl	2	2.69897
44	II	3,5 Br_2	4-CF_3	62.5	1.20412

Figure 1: Common atom of the molecules



N-Benzylsalicylamides (I) N-Benzylsalicylthioamides (II)

Table 2: *k*-Means clustering of compounds using standardized descriptors

Cluster No.	No. of compounds in different clusters	Compounds (Sl nos.) in each clusters																		
1	19	16	17	18	19	22	23	24	27	28	29	34	35	36	37	38	40	41	42	43
2	5	2	11	15	25	44														
3	12	1	3	4	5	6	7	8	9	10	12	13	14							
4	8	20	21	26	30	31	32	33	39											

Stepwise regression

Using stepping criteria based on F value (F = 4.0 for inclusion; F = 3.9 for exclusion), different equations were derived after successive addition of E-state parameters.

$$pC_{14d} = 0.4427(\pm 0.211) + 3.2656(\pm 0.412)S_1$$

$$n_{Training} = 33, R^2 \quad 0.669, R_a^2 \quad 0.658, S \quad 0.454, F \quad 62.68(df\ 1,31),$$

$$Q^2 = 0.6224, PRESS \quad 7.31, n_{Test} \quad 11, R_{pred}^2 \quad 0.7805 \qquad \ldots\ldots\ldots\ldots(1)$$

The standard errors of the respective E-state indices are mentioned within parentheses. Eq. (1) could explain 65.8% of the variance (adjusted coefficient of variation) and leave – one – out predicted variance was found to be 62.24%. While Eq. (1) was applied for prediction of test set compounds, the predictive R^2 value for the test set was found to be 0.7805. The positive coefficient of S_1 indicates that activity increases with increase in E-state value of atom 1. Position 1 indicates the importance of connecting moiety methylcarboxamido / methylthiocarboxamido group between two substituted phenyl groups. Compounds like **17, 18 and 36** with high values of E-state parameter for atom 1 showed comparative higher activity.

Table 3: Observed and calculated antimycobacterial activities from different models

Sl. No.	Obs[a] (pC_{14d})	Cal[b]	Cal[c]
Training set			
1	1.49485	1.430547	1.390504
2	1.20412	0.841944	0.90055
4	1.49485	1.301539	1.180623
5	1.20412	0.672588	0.648611
6	1.49485	1.150524	1.12773
8	0.60206	1.45692	1.445844
9	0.60206	1.371697	1.374208
10	1.20412	1.454824	1.450997
13	1.20412	1.481559	1.420413
14	1.20412	1.399875	1.354519
15	1.20412	0.640296	0.662559
16	2.69897	2.648044	2.62104
18	3.30103	2.55841	2.555367
19	2.39794	2.572284	2.57091
20	2.69897	2.455712	2.51508
22	2.69897	2.645254	2.668074
24	3	2.537397	2.577588
25	2.69897	1.91694	2.074348
26	1.79588	2.354888	2.339839
27	2.09691	2.517688	2.467847
28	2.39794	2.512528	2.426649
30	1.79588	2.076732	2.009422
31	2.09691	2.02243	1.918077
33	1.20412	1.883576	1.903738
34	2.39794	2.361513	2.372596
35	3	2.664374	2.683023
36	3.30103	2.667912	2.69071
38	3	2.665816	2.700589
39	2.39794	2.250529	2.413421
40	1.79588	2.342822	2.241094
41	2.09691	2.438198	2.463413
43	2.69897	2.610867	2.600547
44	1.20412	1.851285	1.917687
3	1.49485	1.295117	1.213466
7	0.90309	1.453382	1.440102
11	1.20412	1.028194	1.133237
12	1.20412	1.18446	1.246312
17	3.30103	2.643632	2.627002
21	2.69897	2.40141	2.423395
23	2.69897	2.371761	2.856792
29	2.09691	2.261114	2.114386
32	1.49485	1.697752	1.782706
37	3	2.582689	2.619074
42	3	2.3355	2.421884

[a] Observed activity (ref. 2); [b] Calculated from eq. (1); [c] Calculated from eq. (2);

Table 4: Statistical comparison of different models

Type of statistical methods	R^2	R_a^2	Q^2	R^2_{pred}
Stepwise regression	0.669	0.658	**0.6224**	0.7805
PLS	0.7106	0.6913	0.606	**0.8057**

*The best values of different parameters are shown in bold.

Table 5: Intercorrelation among descriptors used in equation 1 and 2

	S_1	S_5	S_6	S_8	S_9	S_{13}	S_{14}
S_1	1	0.303	0.945	0.911	-0.895	0.423	0.134
S_5	0.303	1	0.478	0.131	-0.004	0.134	0.094
S_6	0.945	0.478	1	0.805	-0.733	0.461	0.187
S_8	0.911	0.131	0.805	1	-0.930	0.341	0.086
S_9	-0.895	-0.004	-0.733	-0.930	1	-0.143	0.094
S_{13}	0.423	0.134	0.461	0.341	-0.143	1	0.757
S_{14}	0.134	0.094	0.187	0.086	0.094	0.757	1

PLS

The number of optimum components was 2 to obtain the final equation (optimized by cross validation). Based on the standardized regression coefficients, the following variables were selected for the final equation:

$$pC_{14d} = -0.6348 + 0.8473S_1 + 0.1423S_5 + 1.2462S_6$$
$$+0.39435S_8 - 0.04726S_9 + 0.13799S_{13} - 0.0393S_{14}$$
$$n_{Training} = 33, R^2 \quad 0.7106, R_a^2 = 0.6913, Q^2 \quad 0.606, S \quad 0.037$$
$$PRESS = 7.62, F = 69.18(df 1, 31), n_{Test} = 11, R^2_{pred} \quad 0.8057$$
$$\text{.....(2)}$$

Eq. (2) could explain 69.13% of the variance (adjusted coefficient of variation) and leave – one – out predicted variance was found to be 60.60%. While Eq. (2) was applied for prediction of test set compounds, the predictive R^2 value for the test set was found to be 0.8057. The negative coefficients of S_9 and S_{14} indicate that activity decreases with increase in E-state value of atoms 9 and 14 respectively. Compounds with high values of E-state parameter for atom 9 (S_9) (like **1, 3, 5 and 10)** and for atom 14 (S_{14}) (like **10 and 33)** showed comparatively poor activity. The positive coefficient of S_1, S_5, S_6, S_8, and S_{13} indicates that activity increases with increase in E-state value of atom 1, 5, 6, 8 and 13 respectively. Compounds with high values of E-state parameter for atom 5 (S_5) (like **18, 35** and **36)** for atom 6 (S_6) (like **35, 36**, **37 and 38)** for atom 8 (S_8) (like **22 and 23)** and for atom 13 (S_{13}) (like **17 and 38)** showed comparatively higher activity

Conclusions:

The whole dataset (n=44) was divided into a training set (33 compounds) and a test set (11 compounds) based on *k*-means clustering of the standardized descriptor matrix and models were developed from the training set. The predictive ability of the models was judged from the prediction of the activity of the test set compounds. All the developed models indicate the importance of connecting moiety methylcarboxamido / methylthiocarboxamido group between two substituted phenyl groups. From Table 5 it was observed that there is intercorrelation among descriptors like between S_1 and S_6 (r=0.945), between S_1 and S_8 (r=0.911), between S_1 and S_9 (r=0.895). However the both the equations have passed the threshold limit both in internal and external validation.

References

1. Bloch K.C., Zwerling L., Pletcher M.J., Hahn J.A., Gerberding J.L. and Ostroff S.M. Incidence and clinical implications of isolation of Mycobacterium kansassi: results of a 5-year, population-based study. Ann. Int. Med. 1998, 129, 698– 704.
2. Dolezal R., Waisser K., Petrlikova E., Kunes J., Kubicova L., Machacek M., Kaustova J. and Martin Dahse H., N Benzylsalicylthioamides:

Highly active potential Antituberculotics, Arch. Pharm. Chem. Life Sci., 2009, 342, 113-119.

3. SRETSA is statistical software in Visual Basic, developed by Ray S. and Biswas R. and standardized using known data sets.
4. Leonard J.T. and Roy K., On Selection of Training and Test Sets for the Development of Predictive QSAR models, QSAR Comb. Sci., 2006, 25, 235-251.
5. MINITAB is statistical software of Minitab Inc, USA, http://www.minitab.com.
6. Darlington R.B., Regression and linear models, McGraw Hill, New York, 1990.
7. Wold S., PLS for multivariate linear modeling in Van de Waterbeemd H., (Ed.), Chemometric Methods in Molecular Design (Methods and Principles in Medicinal Chemistry), VCH, Weinheim, 1995, 195-218.
8. Fan Y., Shi L.M., Kohn K.W., Pommier Y. and Weinstein J.N., Quantitative structure-antitumor activity relationships of camptothecin analogs: Cluster analysis and genetic algorithm-based studies, J. Med. Chem., 2001, 44, 3254-3263.
9. Snedecor G.W. and Cochran W.G., Statistical Methods, Oxford & IBH Publishing Co. Pvt. Ltd.: New Delhi, 1967.
10. Debnath A.K., in Ghose A.K. and Viswanadhan V.N., (Eds) Combinatorial library design and evaluation: Principles, software tools, and applications in drug discovery, Marcel Dekker, New York, 2001, 73–129.
11. Roy K., On some aspects of validation of predictive QSAR models, Expert Opin. Drug Discov., 2007, 2, 1567-1577.
12. Roy P.P., Leonard J.K. and Roy K., Exploring the impact of the size of training sets for the development of predictive QSAR models, Chemom. Intell. Lab. Sys. 2008, 90, 31-42

*****