Topological modeling of antimycobacterial activity of 3-formyl rifamycin SV derivatives

Omar Deeb,[a,*] Jyoti Singh,[b] R.G. Varma,[c] and Padmakar V. Khadikar[d]

[a]Faculty of Pharmacy, Al-Quds University, P.O. Box 20002, Jerusalem, Palestine
[b]QSAR and Computer Chemical Laboratories, A.P.S. University, Rewa – 486 003, India
[c]Department of Chemistry, PMB Gujarati Science College, Indore 452 010, India
[d]Research Division, Laxmi Fumigation and Pest Control Pvt. Ltd., 3, Khatipura, Indore – 452 007, India
E-mail: deeb2000il@yahoo.com

Abstract

The paper describes topological modeling of antimycobacterial activity of 3-formyl rifamycin SV derivatives using a large series of molecular vis-à-vis topological descriptors. For the set of 53 derivatives of 3-formyl rifamycin SV no one variable model is possible, however, in multiparametric regression excellent model is obtained for modeling the activity. The results are discussed using variety of statistical parameters.

Keywords: Topological modeling, rifamycin, regression analysis, QSAR, topological index

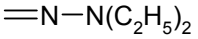
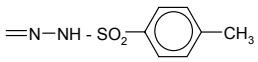
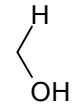
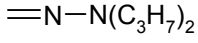
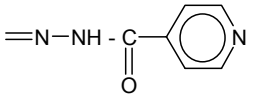
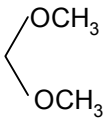
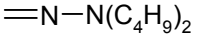
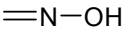
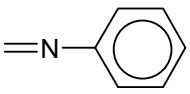
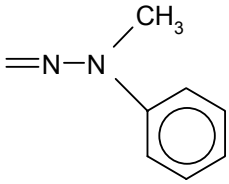
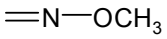
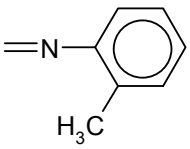
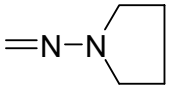
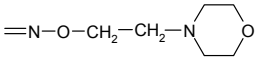
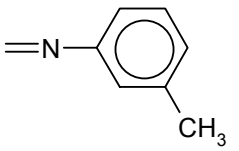
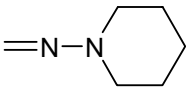
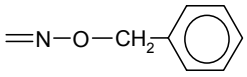
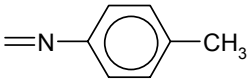
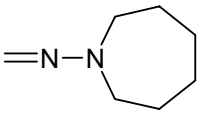
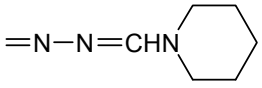
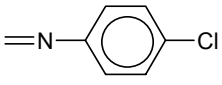
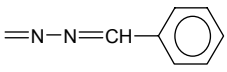
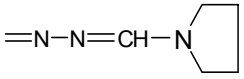
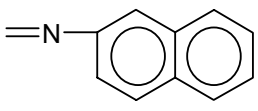
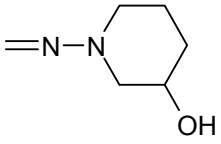
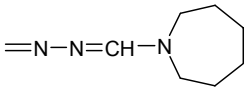
Introduction

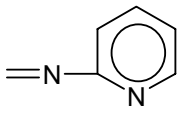
Rifamycins are a group of chemically related antibiotics obtained from *Streptomyces mediterrani.* They belong to a new class of antibiotics that contain a macrocyclic ring bridged across two non-adjacent (ansa) portions of an aromatic nucleus and called ansamycins [1]. The rifamycins and many of their semi-synthetic derivatives have a broad spectrum of anti-microbial activity [1,2]. They are most notably active against gram-positive bacteria and *Mycobacterium tuberculosis*. However, they are also active against some gram-negative bacteria and many viruses. They form a class of antibiotics with a specific potency as drug against tuberculosis via inhibition of the DNA-dependent RNA polymerase [3]. Rifamycin SV (which lacks a C-4 constituent) and the glycolic acid linked at C-3 has antibacterial activity. 3-formyl rifamycin derivatives (Table 1) are one of the classes of anzamycins widely used against infections caused by ordinary bacteria, tuberculosis and leprosy [4]. In the present study the antibacterial potency, $\log(MIC^{RIA} / MIC^{X})$, of this class of compounds (Fig.1, Table 1) against *Mycobacterium tuberculosis* are subjected to a QSAR analysis using a large set of topological indices (Tables 2

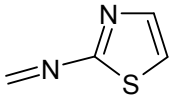
 ©ARKAT USA, Inc.

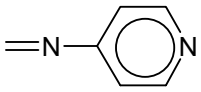
and 3).In these tables as well as in the text this activity is shown as logA. The QSAR modeling is then performed by maximum-R^2 method using step-wise regression analysis [5-7]. The results are discussed below.

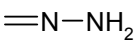
Table 1. Structural details of the compounds used in the present study

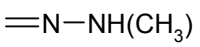
Comp	R	Comp	R	Comp	R
1		19		37	
2		20		38	
3		21		39	
4		22		40	
5		23		41	
6		24		42	
7		25		43	
8		26		44	
9		27		45	

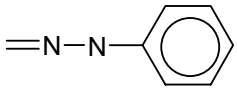
10 

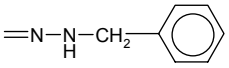
11 

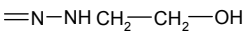
12 

13 

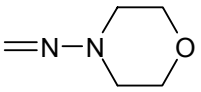
14 

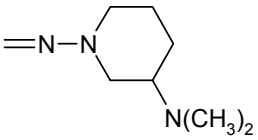
15 

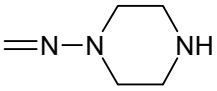
16 

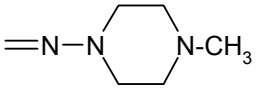
17 

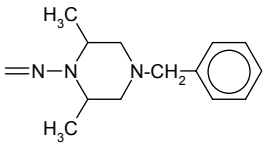
18

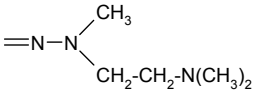
28 

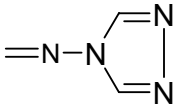
29 

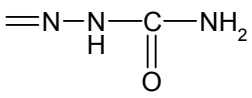
30 

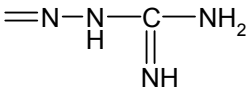
31 

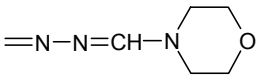
32 

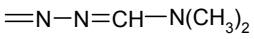
33 

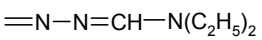
34 

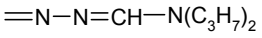
35 

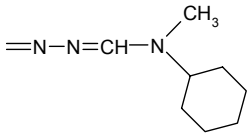
36 

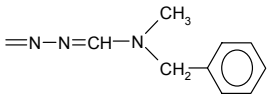
46 

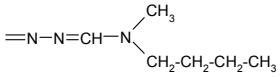
47 

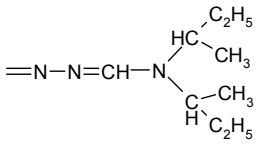
48 

49 

50 

51 

52 

53 

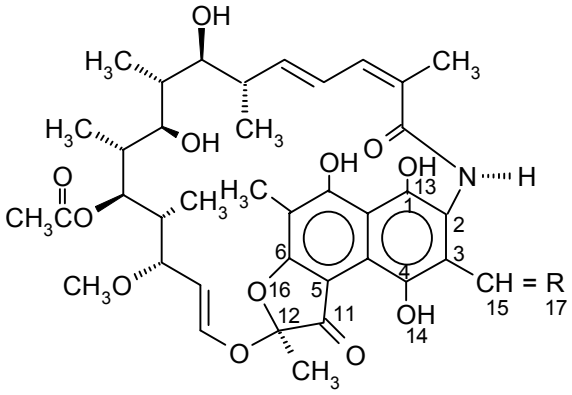


Figure 1. General structure of the compounds used in this study.

Table 2. Topological descriptors used in this study

	Symbol	Definition
1	X0A	Average connectivity index chi-0
2	X5AV	Average valence connectivity index chi-5
3	BAC	Balaban centric index
4	IDDE	Mean information content on the distance degree equality
5	BIC4	Bond information content (neighborhood symmetry of 4-order)
6	LP1	Lovasz-Pelikan index (leading eigen value)
7	EIG1M	Leading eigen value from mass weighted distance matrix
8	SEIGE	Eigen value sum from electronegativity weighted distance matrix
9	VEA1	Eigen vector coefficient sum from adjacency matrix
10	VRE2	Average Randic-type eigen vector-based index from electronegativity weighted distance matrix
11	VEP2	Average eigen vector coefficient sum from polarizability weighted distance matrix
12	VRP1	Randic-type eigen vector-based index from polarizability weighted distance matrix
13	MPC09	Molecular path count of order 09
14	MPC10	Molecular path count of order 10
15	PIPC10	Molecular multiple path count of order 10

Table 3. Observed activity (logA) and topological descriptors used in this study

Comp.	BIC4	LP1	VEA1	X5Av	VRE2	IDDE	MPC09	logA
1	0.88	2.638	4.578	0.022	7.958	5.277	376	0.645
2	0.881	2.638	4.578	0.022	7.977	5.277	376	-1.053
3	0.866	2.644	4.708	0.022	8.232	5.527	423	-1.028
4	0.877	2.639	4.660	0.022	8.591	5.617	457	-1.011
5	0.885	2.639	4.678	0.022	8.672	5.679	466	-1.003

6	0.885	2.639	4.669	0.022	8.678	5.679	464	-1.003
7	0.875	2.639	4.667	0.023	8.685	5.611	465	-1.003
8	0.877	2.639	4.667	0.023	8.681	5.611	465	-0.992
9	0.882	2.639	4.694	0.022	9.003	5.825	505	-0.286
10	0.886	2.639	4.667	0.022	8.675	5.611	465	-1.010
11	0.884	2.639	4.661	0.023	8.516	5.420	453	-1.008
12	0.876	2.639	4.660	0.022	8.589	5.617	457	-0.612
13	0.880	2.639	4.601	0.022	8.058	5.351	391	-0.347
14	0.881	2.639	4.612	0.022	8.153	5.194	405	-1.038
15	0.880	2.639	4.641	0.022	8.683	5.611	445	-0.605
16	0.881	2.639	4.631	0.023	8.786	5.402	439	0.005
17	0.887	2.639	4.620	0.022	8.357	5.401	418	-0.322
18	0.867	2.639	4.630	0.022	8.241	5.563	419	0.970
19	0.863	2.639	4.646	0.022	8.427	5.420	435	1.588
20	0.859	2.639	4.653	0.023	8.614	5.578	445	1.001
21	0.856	2.639	4.657	0.024	8.799	5.689	453	0.015
22	0.877	2.639	4.669	0.022	8.753	5.436	459	1.607
23	0.868	2.639	4.661	0.022	8.522	5.42	453	0.382
24	0.870	2.639	4.660	0.024	8.614	5.617	457	-0.008
25	0.864	2.639	4.660	0.025	8.697	5.578	459	0.397
26	0.879	2.639	4.631	0.022	8.767	5.402	439	-0.297
27	0.888	2.639	4.669	0.024	8.698	5.679	464	0.001
28	0.870	2.639	4.660	0.023	8.609	5.617	457	0.391
29	0.876	2.639	4.678	0.025	8.871	5.800	474	-0.588
30	0.871	2.639	4.660	0.023	8.611	5.617	457	-0.610
31	0.869	2.639	4.667	0.024	8.707	5.611	465	0.000
32	0.865	2.640	4.734	0.025	9.336	5.772	523	0.449
33	0.873	2.639	4.644	0.023	8.734	5.469	440	0.001
34	0.876	2.639	4.661	0.022	8.506	5.420	453	-1.017
35	0.884	2.639	4.626	0.022	8.333	5.365	421	-1.022
36	0.885	2.639	4.626	0.022	8.334	5.365	421	-1.022
37	0.878	2.639	4.671	0.024	8.901	5.761	455	-0.264
38	0.884	2.639	4.465	0.022	8.845	5.800	447	-0.988
39	0.882	2.639	4.601	0.022	8.057	5.351	391	-1.046
40	0.879	2.639	4.612	0.022	8.147	5.194	405	0.032
41	0.874	2.639	4.626	0.023	8.896	5.800	436	0.017
42	0.879	2.639	4.631	0.022	8.777	5.402	439	-0.011
43	0.872	2.639	4.631	0.024	8.789	5.402	439	0.006
44	0.871	2.639	4.631	0.023	8.699	5.611	437	-0.001
45	0.867	2.639	4.631	0.025	8.869	5.767	441	-0.589

46	0.872	2.639	4.631	0.023	8.784	5.402	439	-0.294
47	0.870	2.639	4.624	0.022	8.426	5.456	423	-0.316
48	0.866	2.639	4.628	0.022	8.611	5.611	431	-0.301
49	0.862	2.639	4.629	0.023	8.791	5.767	435	-0.286
50	0.874	2.639	4.633	0.025	8.956	5.793	441	0.016
51	0.881	2.639	4.630	0.023	9.018	5.680	439	-0.576
52	0.884	2.639	4.627	0.023	8.726	5.469	430	-0.595
53	0.857	2.639	4.637	0.023	8.929	5.616	443	-0.574

Results and Discussion

So far, no QSAR studies with rifamycins employing molecular descriptors mentioned in Table 3 were used to quantify and elucidate potentially relevant chemical reactivity patterns of the drugs. The literature data on the antibacterial potential of 3-formylrifamycin SV derivatives (Table 3) was used for preparing models with excellent statistics. The correlation of the antibacterial activity with the molecular descriptors used is given in Table 4.

A preliminary regression analysis (Table 5) has indicated that none of the molecular descriptors used singly is capable of modeling the activity. However, the data presented in Table 5 did show that the variable BIC4 (Bond information content, neighborhood symmetry of 4-order) is the promising descriptor to be used in multiparametric regression analysis. It means that multiparametric model(s) will invariably contain this BIC4 as one of the correlating parameters.

Before a multivariate analysis is undertaken it is convenient to normalize the data in certain ways in order to make the detection of significant correlations easier. Normally, it is sufficient to preprocess the data by means of auto-scaling and mean-centering the variables. Auto-scaling gives each variable unit variance and hence the same chance to contribute to a estimated model, while mean-scaling facilitates interpretation. This can be achieved by obtaining correlation matrix. Such a correlation matrix, as stated earlier, is presented in Table 4. An examination of the correlation matrix (Table 4) shows molecular descriptors used did exhibit linear correlation. That is, model containing such descriptors will suffer from the defect due to collinearity, which statistically is not allowed. Such cases will be examined using Randic [8] recommendations discussed in the following section.

Table 4. Correlation matrix for the activity and the descriptors used in this study

	logA	X0A	X5AV	BAC	IDDE	BIC4	LP1	EIG1M
logA	1							
X0A	0.02475	1						
X5AV	0.14924	-0.33202	1					
BAC	0.08723	0.74611	0.02572	1				
IDDE	-0.02787	-0.49132	0.53572	0.03119	1			
BIC4	-0.48497	-0.11430	-0.41367	-0.39190	-0.28160	1		
LP1	-0.12101	0.08337	0.00468	0.18069	0.10776	-0.22259	1	
EIG1M	0.13101	-0.55892	0.59003	0.07632	0.73316	-0.28776	0.00823	1
SEIGE	-0.14140	0.06136	0.10853	0.12272	0.09245	0.06470	0.12629	0.13724
MPC09	0.05653	-0.72254	0.41573	-0.17874	0.68451	-0.17186	0.11975	0.68341
MPC10	0.04380	-0.73795	0.43449	-0.18817	0.69658	-0.15014	0.08270	0.70535
PIPC10	-0.23741	-0.61839	0.07881	-0.20740	0.54719	0.09777	0.00523	0.49300
VEA1	0.02565	-0.50512	0.31264	-0.04970	0.57028	-0.22145	0.50043	0.45412
VRP1	0.11257	-0.62598	0.57586	0.02053	0.76045	-0.26281	0.02914	0.98667
VEP2	-0.11631	0.60096	-0.54522	-0.04721	-0.75140	0.26808	-0.02621	-0.97364
VRE2	0.09404	-0.69116	0.57202	-0.06009	0.76779	-0.22689	0.01303	0.97269
	SEIGE	MPC09	MPC10	PIPC10	VEA1	VRP1	VEP2	VRE2
SEIGE	1							
MPC09	-0.13777	1						
MPC10	-0.11535	0.99600	1					
PIPC10	-0.19410	0.80463	0.79886	1				
VEA1	-0.11275	0.88633	0.86689	0.65296	1			
VRP1	0.14575	0.76749	0.78848	0.59164	0.54933	1		
VEP2	-0.17184	-0.76815	-0.78914	-0.60826	-0.55057	-0.99212	1	
VRE2	0.10937	0.80628	0.82737	0.63707	0.58025	0.99511	-0.98697	1

Following maximum-R^2 method [5-7], and using a large set of 15 descriptors and the entire set of 53 compounds, we obtained several models containing 1 to 10 correlating parameters (Table 5) and observed that the models contains compounds **27, 31, 33, 43** and **44** as outliers. The deletion of these compounds gave better results (Table 6). A perusal of Table 6 shows that statistically better models start from model-25. A detailed analysis of these models (Table 7) indicates that they contain one or more correlating parameters in that the coefficient of the correlating parameter is significantly smaller than their respective standard deviation. Such models are not allowed statistically. The deletion of such parameters from the models yielded improved models 33-38 as presented in Table 8. Hence, our further discussion will be centered on these six models: 33-38.The data presented in Tables 8 and 9 indicate that the model 38 is the best model for modeling the antibacterial activity.

It is interesting to mention that the model 38 contains 12 correlating parameters. It becomes necessary to examine the model 38 by applying the rule of thumb [9,10] and searching optimum descriptors that can be used in proposing the models for the data set of 53 (reduced to 48 after

removing five outliers) compounds used in the present study. The limitations and some common pitfalls of multiple regression analysis were pointed out by Tute [9,10]. According to him, there must be a sufficient number of compounds included in the analysis to enable statistical significance to be reached, despite inevitable errors in measurement. A rule of thumb evolved by Tute [9,10] is that the number compounds to be used should be at least three times the number of parameters under consideration. Looking to the data set (48 compounds) and in accordance with the rule of thumb the proposed 12 parametric model is quite justified. In order to confirm this finding we have investigated optimum number of parameters that could be used for modeling the activity of 48 compounds. This we did by plotting graphs between the number of variables and the corresponding R^2 and R^2_A values [Table 10] plotted on the same graph. In our case both R^2 and R^2_A go on increasing with the number of variables and becomes almost constant at 12 parameters. This finding is, therefore, consistent with the results obtained by applying the rule of thumb [9, 10]. Further confirmation is made by calculating the activities from each of the proposed models and comparing them with the experimental (observed) activities. Such comparisons are given in Table 11 and demonstrated in Figures 2-7. The results are in favor of 12 parametric model 38. We have also used the data from Figures 2-7 and obtained correlations between observed and estimated antibacterial activity. This demonstrated by models 39-44 (Table 12), which finally confirmed that the proposed 12 parametric model 38 is the most appropriate model for modeling the activity.

Table 5. Model summary considering all the 53 Compounds

Model	R	R^2	Adjusted R^2	Std. error of the Estimate	Change Statistics				
					R^2 Change	F Change	df1	df2	Sig. F Change
1	0.481	0.231	0.216	0.68767	0.231	16.314	1	61	0.000
2	0.519	0.269	0.240	0.57865	0.038	2.602	1	50	0.113
3	0.581	0.338	0.297	0.55626	0.069	5.106	1	49	0.028
4	0.613	0.375	0.323	0.54592	0.037	2.873	1	48	0.097
5	0.645	0.416	0.353	0.53363	0.040	3.237	1	47	0.078
6	0.664	0.441	0.368	0.52741	0.026	2.116	1	46	0.153
7	0.683	0.467	0.384	0.52088	0.026	2.161	1	45	0.149
8	0.693	0.480	0.386	0.52017	0.013	1.122	1	44	0.295
9	0.719	0.517	0.416	0.50716	0.037	3.286	1	43	0.077
10	0.712	0.506	0.417	0.50684	-0.011	0.945	1	45	0.336
11	0.764	0.584	0.497	0.47082	0.077	7.991	1	43	0.007
12	0.763	0.582	0.606	0.46621	-0.001	0.143	1	45	0.707
13	0.787	0.620	0.541	0.44985	0.038	4.259	1	43	0.045
14	0.797	0.635	0.548	0.44639	0.015	1.670	1	42	0.203

Table 6. Model summary for the set of 48 compounds after deleting five compounds(26,31,33, 43 and 44) as outliers

Model	R	R^2	Adjusted R^2	Std. error of the estimate	Change Statistics				
					R^2 Change	F Change	df1	df2	Sig. F Change
15	0.480	0.235	0.219	0.61086	0.235	14.146	1	46	0.000
16	0.639	0.290	0.259	0.69492	0.055	3.948	1	45	0.068
17	0.570	0.324	0.278	0.58705	0.034	2.216	1	44	0.144
18	0.633	0.400	0.345	0.55940	0.076	4.457	1	43	0.024
19	0.655	0.429	0.362	0.55215	0.029	2.136	1	42	0.151
20	0.684	0.468	0.391	0.53942	0.039	3.006	1	41	0.090
21	0.702	0.493	0.404	0.63355	0.024	1.906	1	40	0.175
22	0.720	0.518	0.419	0.52669	0.025	2.049	1	39	0.160
23	0.746	0.556	0.451	0.51193	0.038	3.281	1	38	0.078
24	0.769	0.591	0.481	0.49782	0.035	3.186	1	37	0.082
25	0.792	0.627	0.513	0.48203	0.036	3.464	1	36	0.071
26	0.809	0.655	0.536	0.47065	0.027	2.761	1	35	0.106
27	0.805	0.649	0.541	0.46797	-0.006	0.591	1	37	0.447
28	0.812	0.660	0.543	0.46705	0.011	1.142	1	35	0.283
29	0.810	0.656	0.551	0.46281	-0.003	0.350	1	37	0.558
30	0.820	0.673	0.561	0.45795	0.017	1.770	1	35	0.192
31	0.866	0.750	0.654	0.40644	0.077	10.433	1	34	0.003
32	0.865	0.748	0.661	0.40237	-0.002	0.303	1	36	0.586

Table 7. Details of the statistically significant models 25-32

Regression Equation for Model No.25 :

logA = 2150.474 (± 665.214) – 29.953 (± 11.254) BIC4 – 885.635 (± 279.172) LP1
– 0.001 (± 0.000) PIPC10 + 38.664 (± 17.006) VEA1
– 251.831 (± 108.540) X5AV + 8.836 (± 2.766) VRE2 – 0.745 (± 0.703) IDDE – 0.097 (± 0.032) MPC10 + 8.712E-02 (± 0.039) MPC09
– 0.036 (± 0.015) EIG1M – 0.993 (± 0.534) SEIGE

Regression Equation for Model No.26 :

logA = 2056.90 (± 651.954) – 20.756 (± 12.303) BIC4 – 825.621 (± 274.967) LP1

– 0.001 (± 0.000) PIPC10 + 34.753 (± 16.771) VEA1
– 208.754 (± 109.104) X5AV + 7.115 (± 2.893) VRE2 – 0.536 (± 0.698) IDDE – 104 (± 0.032) MPC10 + 0.100 (± 0.039) MPC09 – 0.048 (± 0.016) EIG1M
– 1.553 (± 0.620) SEIGE – 287.448 (± 172.991) VEP2

Regression Equation for Model No.27 :
logA = 2012.690 (± 645.707) – 18.799 (± 11.968) BIC4 – 805.358 (± 272.139) LP1
– 0.001 (± 0.000) PIPC10 + 32.763 (± 16.475) VEA1
– 223.505 (± 106.790) X5AV + 7.027 (± 2.874) VRE2
– 0.107 (± 0.031) MPC10 + 0.105 (± 0.038) MPC09 – 0.050 (± 0.016) EIG1M
– 1.649 (± 0.604) SEIGE – 311.397 (± 169.190) VEP2

Regression Equation for Model No.28 :
logA = 1737.177 (± 694.112) – 15.240 (± 12.401) BIC4 – 682.864 (± 294.809) LP1
– 0.001 (± 0.000) PIPC10 + 22.802 (± 18.902) VEA1
-210.345 (± 107.289) X5AV + 2.862 (± 4.840) VRE2 - 0.103 (± 0.031) MPC10 + 0.111 (± 0.038) MPC09 – 0.070 (± 0.024) EIG1M – 1.984 (± 0.680) SEIGE – 242.243 (± 180.837) VEP2 + 5.150E-02 (± 0.048) VRP1

Regression Equation for Model No.29 :
logA = 1561.558 (± 621.704) – 12.979 (± 11.690) BIC4 – 602.395 (± 259.165) LP1
– 0.001 (± 0.000) PIPC10 + 16.683 (± 15.675) VEA1
-193.821 (± 102.648) X5AV + 0.097 (± 0.029) MPC10
+ 0.112 (± 0.038) MPC09 – 0.075 (± 0.023) EIG1M – 2.109 (± 0.640) SEIGE – 232.542 (± 178.458) VEP2 + 7.446E-02 (± 0.028) VRP1

Regression Equation for Model No.30 :
logA = 2168.174 (± 765.755) – 18.004 (± 12.168) BIC4 – 854.825 (± 319.015) LP1
– 0.001 (± 0.000) PIPC10 + 35.663 (± 21.075) VEA1
-200.192 (± 101.681) X5AV + 0.105 (± 0.030) MPC10
+ 0.101 (± 0.039) MPC09 – 0.057 (± 0.026) EIG1M – 1.922 (± 0.649) SEIGE – 319.790 (± 188.368) VEP2 + 5.379E-02 (± 0.032) VRP1
- 0.004 (± 0.003) BAC

Regression Equation for Model No.31 :
logA = 1565.801 (± 704.755) – 25.565 (± 11.050) BIC4 – 753.356 (± 284.874) LP1
– 0.001 (± 0.000) PIPC10 + 28.792 (± 18.825) VEA1
-283.412 (± 93.851) X5AV + 0.087 (± 0.027) MPC10
+ 0.106 (± 0.034) MPC09 – 0.060 (± 0.023) EIG1M – 1.804 (± 0.577) SEIGE – 99.386 (± 180.573) VEP2 + 9.821E-02 (± 0.032) VRP1

- 0.045 (± 0.013) BAC + 440.470 (± 136.371) X0A

Regression Equation for Model No.32 :

logA = 1445.987 (± 663.583) – 27.100 (± 10.586) BIC4 – 717.297 (± 274.463) LP1
– 0.001 (± 0.000) PIPC10 + 25.844 (± 17.866) VEA1
– 298.788 (± 88.700) X5AV + 0.083 (± 0.026) MPC10
+ 0.106 (± 0.034) MPC09 – 0.063 (± 0.023) EIG1M – 1.734 (± 0.557) SEIGE + 0.109 (± 0.024) VRP1 - 0.047 (± 0.012) BAC + 468.834 (± 124.995) X0A

Table 8. Regression equations for the improved models(33-38)

Regression Equation for Model No.33 :

logA = 2049.932 (± 698.808) – 37.840 (± 11.254) BIC4 – 844.769 (± 293.312) LP1
– 0.001 (± 0.000) PIPC10 + 42.459 (± 17.817) VEA1
– 252.244 (± 114.282) X5AV + 7.304 (± 2.822) VRE2
– 0.036 (± 0.018) MPC10 – 0.028 (± 0.016) EIG1M – 1.059 (± 0.561) SEIGE
- 0.950 (± 0.734) IDDE

Regression Equation for Model No.34 :

logA = 2012.690 (± 645.707) – 18.799 (± 11.968) BIC4 – 805.358 (± 272.139) LP1
– 0.001 (± 0.000) PIPC10 + 32.763 (± 16.475) VEA1
– 223.505 (± 106.790) X5AV + 7.027 (± 2.874) VRE2
– 0.107 (± 0.031) MPC10 + 0.050 (± 0.016) EIG1M – 1.649 (± 0.604) SEIGE + 0.105 (± 0.038) MPC09 - 311.397 (± 169.190) VEP2

Regression Equation for Model No.35 :

logA = 1561.558 (± 621.704) – 12.979 (± 11.690) BIC4 – 602.395 (± 259.165) LP1
– 0.001 (± 0.000) PIPC10 + 16.683 (± 15.675) VEA1
– 193.821 (± 102.648) X5AV – 0.097 (± 0.029) MPC10
- 0.075 (± 0.023) EIG1M – 2.109 (± 0.640) SEIGE + 0.112 (± 0.038) MPC09
– 232.542 (± 178.458) VEP2 + 0.074 (± 0.028) VRP1

Regression Equation for Model No.36 :

logA = 2168.174 (± 765.755) – 18.004 (± 12.168) BIC4 – 854.825 (± 319.015) LP1
- 0.001 (± 0.000) PIPC10 + 35.663 (± 21.075) VEA1
- 200.192 (± 101.681) X5AV - 0.105 (± 0.030) MPC10
- 0.057 (± 0.026)EIG1M – 1.922 (± 0.649) SEIGE + 0.101 (± 0.039) MPC09
- 319.790 (± 188.368) VEP2 + 0.054 (± 0.032) VRP1 – 0.004 (± 0.003) BAC

Regression Equation for Model No.37 :

logA = 2585.068 (± 742.443) – 21.696 (± 12.226) BIC4 – 1021.705 (± 310.696) LP1
- 0.001 (± 0.000) PIPC10 + 49.688 (± 19.831) VEA1

- 179.361 (± 103.439) X5AV + 0.099 (± 0.030) MPC10
- 0.018 (± 0.012) EIG1M – 1.459 (± 0.602) SEIGE + 0.082 (± 0.038) MPC09
- 498.307 (± 159.364) VEP2 – 0.006 (± 0.003) BAC

Regression Equation for Model No.38 :

logA = 1445.987 (± 663.583) – 27.100 (± 10.586) BIC4 – 717.297 (± 274.463) LP1
- 0.001 (± 0.000) PIPC10 + 25.844 (± 17.866) VEA1
- 298.788 (± 88.700) X5AV + 0.083 (± 0.026) MPC10
- 0.063 (± 0.023) EIG1M – 1.734 (± 0.557) SEIGE + 0.106 (± 0.034) MPC09
- 0.047 (± 0.012) BAC + 0.109 (± 0.024) VRP1 + 468.834 (± 124.995) X0A

Table 9. Regression parameters for the proposed models

Model	R	R^2	Adjusted R^2	Standard Error of Estimate (S.E.)	F	Q = R/S.E.
33	0.759	0.575	0.461	0.5075	5.013	1.4955
34	0.805	0.649	0.541	0.4680	6.044	1.7202
35	0.810	0.656	0.551	0.4628	6.253	1.7502
36	0.820	0.673	0.561	0.4579	6.002	1.7906
37	0.804	0.647	0.539	0.4694	5.989	1.7129
38	0.865	0.748	0.661	0.4024	8.635	2.1497

Table10. Number of variables, R^2 and R^2_A used in deciding optimum number of descriptors

Variables used	R^2	R^2_A
1	0.235	0.219
2	0.290	0.259
3	0.324	0.278
4	0.400	0.345
5	0.429	0.362
6	0.468	0.391
7	0.493	0.404
8	0.518	0.419
9	0.556	0.451
10	0.591	0.481
11	0.627	0.513
12	0.655	0.536
11	0.649	0.541
12	0.660	0.543
11	0.656	0.551

12	0.673	0.561
13	0.750	0.654
12	0.748	0.661

Table 11. Observed and estimated activity for the proposed models (33-38)

Comp.	logA (Obs)	logA (Est) 33	logA (Est) 34	logA (Est) 35	logA (Est) 36	logA (Est) 37	logA (Est) 38
1.	0.645	0.6983	0.72934	0.6598	0.66512	0.68512	0.47734
2.	-1.053	-0.20213	-0.59346	-0.66626	-0.61027	-058658	-0.64871
3.	-1.028	-0.83072	-0.84273	-0.90617	-0.8766	-0.87171	-0.96761
4.	-1.011	-0.5853	-0.86775	-0.88153	-0.84843	-0.86962	-0.57546
5.	-1.003	-0.64038	-0.79426	-0.86517	-0.77143	-0.60266	-0.89839
6.	-1.003	-0.83974	-0.84849	-0.88785	-0.89443	-0.8019	-1.03004
7.	-1.003	-0.63215	-0.62535	-0.74068	-0.71834	-0.58635	-0.8766
8.	-0.992	-0.91025	-0.95011	-1.09043	-1.0665	-0.90482	-1.2021
9.	-0.286	-0.95277	-0.33378	-0.18639	-0.27862	-0.49604	-0.27284
10.	-1.01	-1.15035	-1.09711	-1.17573	-1.19748	-1.10013	-1.2624
11.	-1.008	-0.56567	-0.64694	-0.73464	-0.72442	-0.79337	-1.08736
12.	-0.612	-0.70379	-1.08209	-1.12943	-1.0743	-1.0488	-0.75199
13.	-0.347	-0.74515	-0.98397	-0.91378	-0.95117	-1.02309	-0.85368
14.	-1.038	-0.44434	-0.214	-0.12998	-0.15134	-0.19002	-0.19878
15.	-0.605	-0.49053	-0.59073	-0.47531	-0.53221	-0.63888	-0.70262
16.	0.005	-0.09461	-0.0474	-0.14753	-0.12338	-0.14877	-0.27945
17.	-0.322	-0.9797	-0.63008	-0.62608	-0.65524	-0.54909	-0.28776
18.	0.97	0.18356	0.73028	0.73794	0.71731	0.6067	0.7363
19.	1.588	0.63322	0.77012	0.74005	0.76694	0.82206	0.95819
20.	1.001	0.62532	0.65065	0.59871	0.61832	0.69457	0.58878
21.	0.015	0.65595	0.74401	0.59936	0.67351	0.8352	0.84228
22.	1.607	0.58051	0.43541	0.50786	0.57385	0.45962	0.80781
23.	0.382	0.73955	0.79989	0.64863	0.76606	0.78066	0.42488
24.	-0.008	-0.08723	0.09755	0.07836	0.09753	0.09904	0.09681
25.	0.397	0.01175	-0.00057	0.02618	0.06088	0.07394	-0.0158
26.	-0.297	-0.27448	-0.41926	-0.43545	-0.38549	-0.44121	-0.44103
27.	-	-	-	-	-	-	-
28.	0.391	-0.11391	-0.08724	-0.12858	-0.08851	-0.07055	0.11126
29.	-0.588	-0.41253	-0.29328	-0.14231	-0.36262	-0.50487	-0.32441
30.	-0.61	-0.03171	0.07028	0.03	0.05953	0.05963	0.20386
31.	-	-	-	-	-	-	-
32.	0.449	0.36676	0.06654	0.24138	0.15484	0.17411	0.38368
33.	-	-	-	-	-	-	-

34.	-1.017	-0.27581	-0.34813	-0.44782	-0.30265	-0.18615	-0.63547
35.	-1.022	-1.16166	-1.1392	-1.06745	-1.17981	-1.06618	-1.32752
36.	-1.022	-1.08147	-0.97931	-0.87921	-1.00977	-0.93264	-1.20022
37.	-0.264	0.01082	-0.22455	-0.05294	-0.10541	-0.42704	-0.01565
38.	-0.988	-0.64675	-0.68913	-0.76226	-0.74796	-0.7201	-0.76913
39.	-1.046	-0.93652	-1.19611	-1.09209	-1.13968	-1.22023	-0.99449
40.	0.032	-0.5074	-0.36216	-0.14016	-0.18241	-0.29181	-0.08311
41.	0.017	-0.35445	0.0051	-0.28899	-0.03924	0.21013	-0.23521
42.	-0.011	0.13914	0.12489	0.26907	0.22127	-0.02892	0.36617
43.	-	-	-	-	-	-	-
44.	-	-	-	-	-	-	-
45.	-0.589	-0.58476	-0.4968	-0.49077	-0.40055	-0.40308	-0.53955
46.	-0.294	-0.29133	-0.67308	-0.69966	-0.64537	-0.69448	-0.44533
47.	-0.316	-0.48595	-0.49623	-0.34995	-0.4647	-0.49484	-0.10756
48.	-0.301	-0.17517	-0.09435	0.00193	-0.10431	-0.17468	-0.05073
49.	-0.286	-0.25785	-0.35079	-0.3362	-0.37473	-0.31426	-0.24361
50.	0.016	-0.62543	-0.4771	-0.55976	-0.49574	-0.47075	-0.49656
51.	-0.576	-0.32439	-0.23805	-0.29729	-0.16106	-0.11348	-0.25358
52.	-0.595	-0.78749	-0.80581	-0.8893	-0.93242	-0.80746	-0.54265
53.	-0.574	0.32902	0.08534	0.26789	0.01155	-0.13626	-0.58995

'Obs' refers to 'Observed', 'Est' refers to 'Estimated'
The number to the right of (Est) shows the model number.

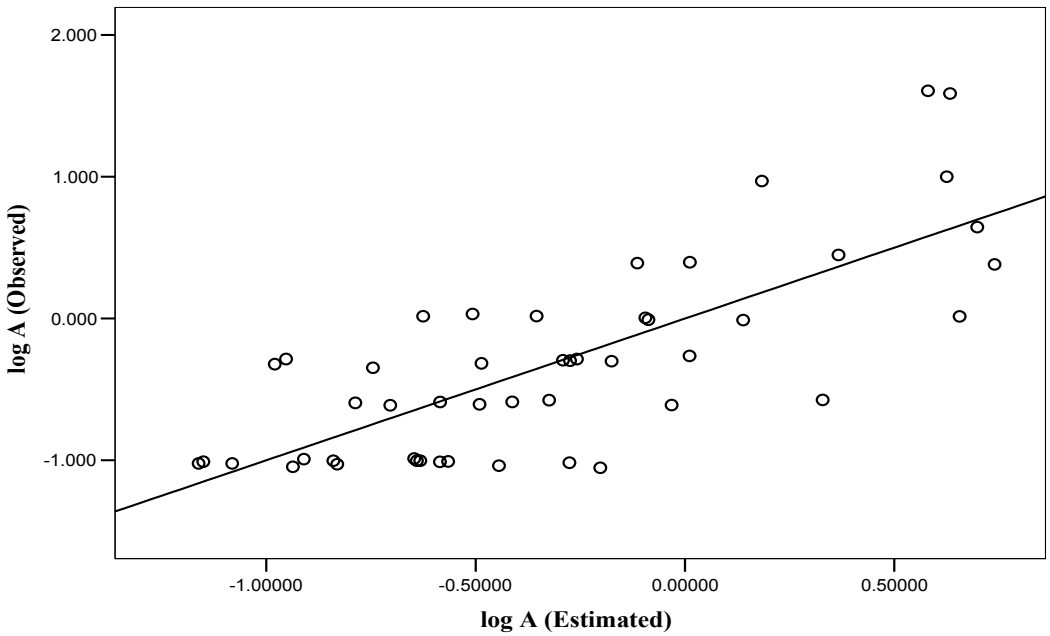


Figure 2. Observed versus estimated activity for model 33.

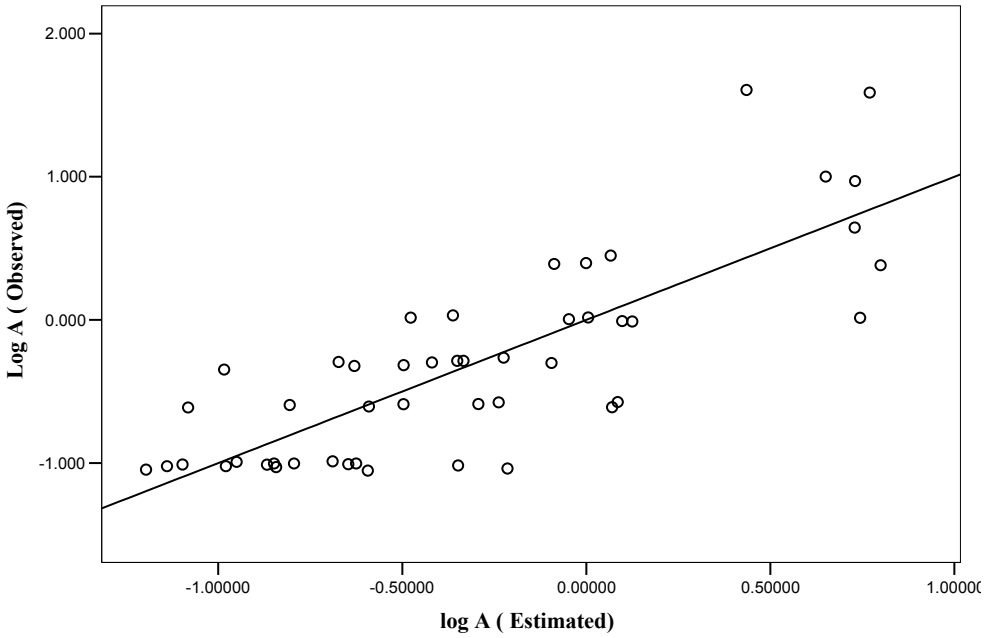


Figure 3. Observed versus estimated activity for model 34.

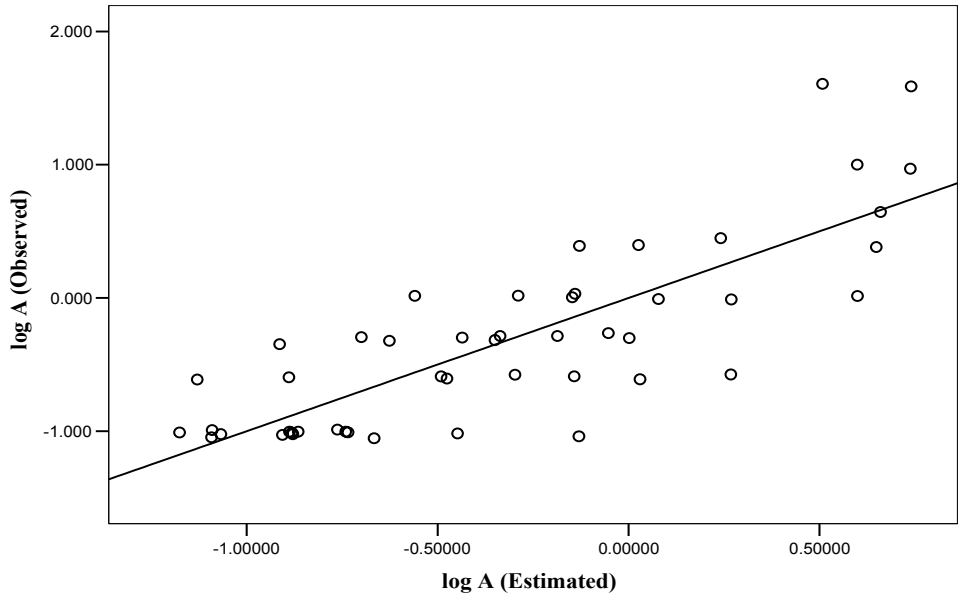


Figure 4. Observed versus estimated activity for model 35.

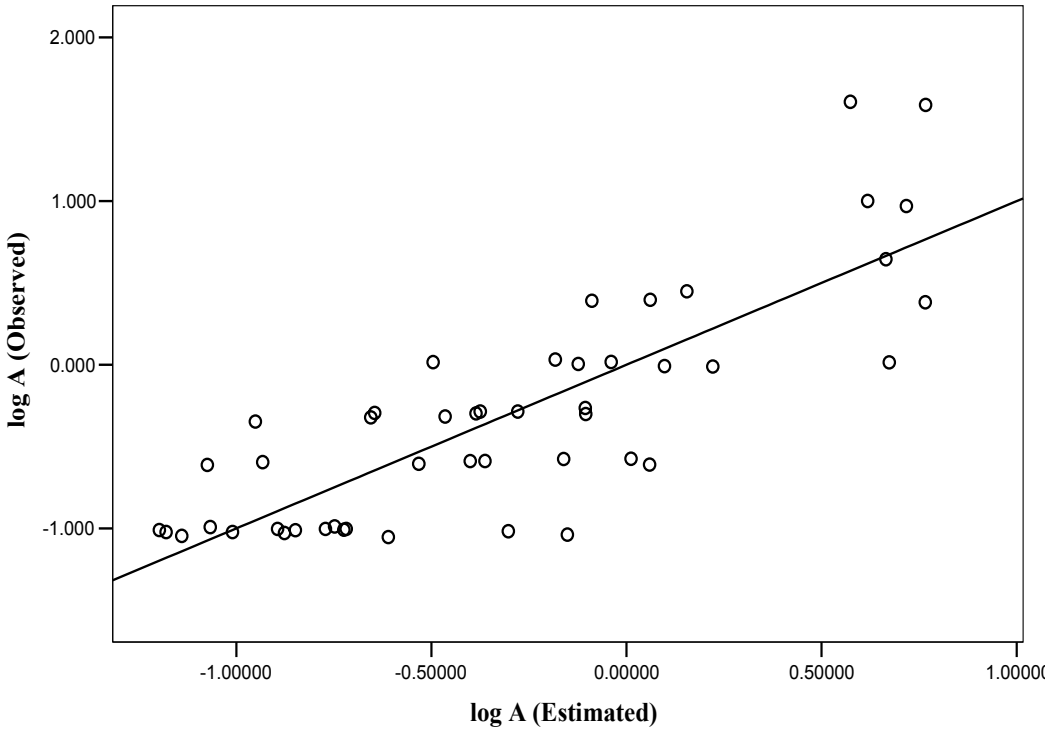


Figure 5. Observed versus estimated activity for model 36.

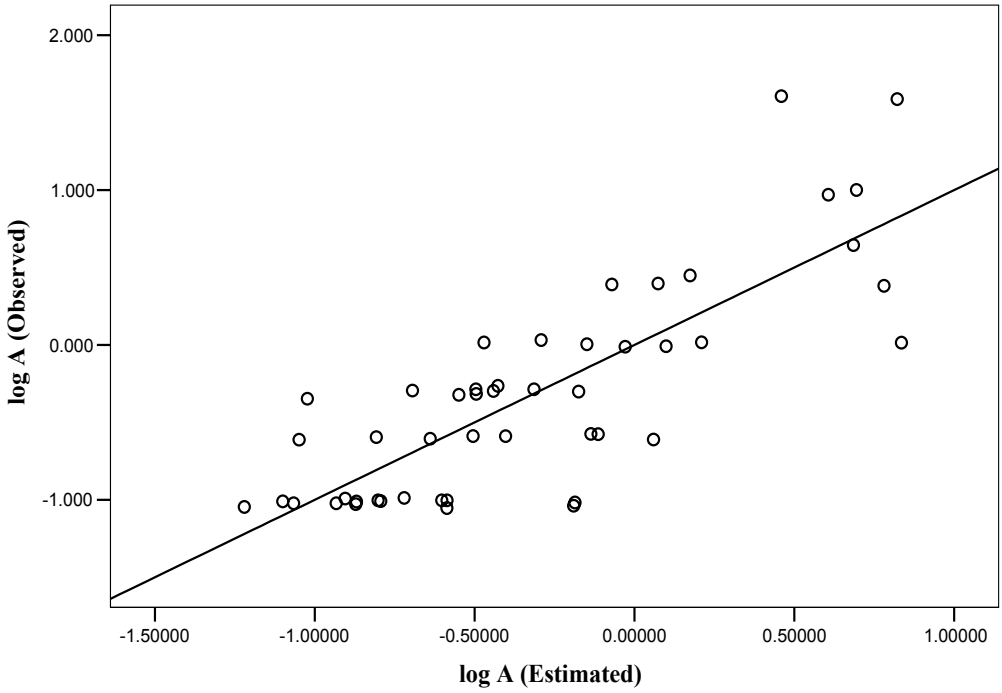


Figure 6. Observed versus estimated activity for model 37.

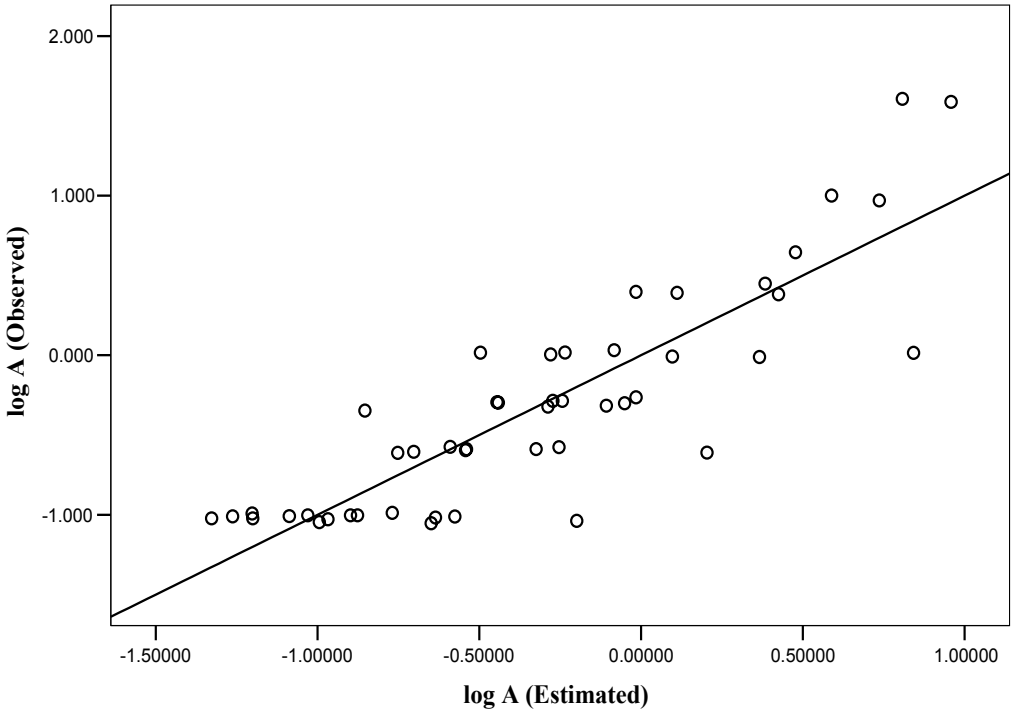


Figure 7. Observed versus estimated activity for model 38.

Table 12. Regression between observed and estimated activity for the proposed models

Model 39 :

$\log A_{(Observed)} = 7.20 \times 10^{-13} (\pm 0.077) + 1.0 (\pm 0.127) \log A_{(Predicted)\ 33}$

n = 48, R = 0.759, $R^2 = 0.575$, $R^2_A = 0.566$, S.E. = 0.455177, F = 62.325

Model 40 :

$\log A_{(Observed)} = 9.83 \times 10^{-13} (\pm 0.069) + 1.0 (\pm 0.108) \log A_{(Predicted)\ 34}$

n = 48, R = 0.805, $R^2 = 0.649$, $R^2_A = 0.641$, S.E. = 0.413989, F = 84.952

Model 41 :

$\log A_{(Observed)} = 7.25 \times 10^{-13} (\pm 0.068) + 1.0 (\pm 0.107) \log A_{(Predicted)\ 35}$

n = 48, R = 0.810, $R^2 = 0.656$, $R^2_A = 0.649$, S.E. = 0.409429, F = 87.885

Model 42 :

$\log A_{(Observed)} = 5.10 \times 10^{-13} (\pm 0.066) + 1.0 (\pm 0.103) \log A_{(Predicted)}36$

n = 48, R = 0.820, $R^2 = 0.673$, $R^2_A = 0.666$, S.E. = 0.399456, F = 94.654

Model 43 :

$\log A_{(Observed)} = 8.18 \times 10^{-13} (\pm 0.069) + 1.0 (\pm 0.109) \log A_{(Predicted)}\ 37$

n = 48, R = 0.804, $R^2 = 0.647$, $R^2_A = 0.639$, S.E. = 0.415228, F = 84.172

Model 44 :

$\log A_{(Observed)} = 1.01 \times 10^{-12} (\pm 0.057) + 1.0 (\pm 0.086) \log A_{(Predicted)\ 38}$

n = 48, R = 0.865, $R^2 = 0.748$, $R^2_A = 0.742$, S.E. = 0.350981, F = 136.189

It is worthy to mention that the aforementioned results and discussions are enough to establish the goodness of fit, but none establishes the goodness of prediction. The proposed models should be excellent model only when they have both excellent fit and excellent predictive power. The latter is now investigated by using the method of cross-validation [5-7]. The various cross-validated parameters estimated for the models 33-38 are given in Table 13. All the cross-validated parameters, except S_{PRESS}, are in favor of the proposed model. From Table 9 and 13 we observed that Se is the same as S_{PRESS} and thus the latter parameter cannot be used in deciding the uncertainty of prediction. In such cases the uncertainty in prediction is judged from yet another cross-validated parameter viz., PSE. The lowest value of PSE decides the uncertainty of prediction. Needless to state the PSE is smallest for the proposed model. Hence, we can conclude that the proposed model 38 has significant fit and predictive power.

Further examination of the data presented in Table 13 indicates that in all the six models PRESS < SSY indicating that these models predict better than chance and thus they can be considered statistically significant. Furthermore, the ratio PRESS / SSY for the model 38 is smaller than 0.4 (0.3378) indicating it to be reasonable QSAR model.

Table 13. Cross validation parameters for the proposed models

Model	PRESS	S_{PRESS}	SSY	R^2_{cv}	PRESS/SSY	PSE
33	9.5305	0.5075	12.9130	0.2619	0.7381	0.4456
34	7.8838	0.4680	14.5597	0.4585	0.5415	0.4053
35	7.7111	0.4628	14.7324	0.4766	0.5234	0.4008
36	7.3400	0.4579	15.1035	0.5140	0.4860	0.3910
37	7.9311	0.4694	14.5124	0.4535	0.5465	0.4065
38	5.6667	0.4024	16.7768	0.6622	0.3378	0.3436

At this stage, it is interesting to comment on R^2_A, which accounts for the adjacent of R^2. It is a measure of the % explained variation in the dependent variable that takes into account the relationship between the number of cases and the number of independent variables in the regression model. Whereas, R^2 will always increase when an independent variable is added. R^2_A will decrease if the added variable doesn't reduce the unexplained variation enough to offset the loss of degrees of freedom. If a variable is added that does not contribute its fair share, the R^2_A will actually decline. A perusal of Table 10 shows that as we pass from a ten parametric model to 12-parametric model, R^2_A go on increasing indicating that in each case the added parameter has enough contribution to the proposed model.

From the data presented in Table 8 we observed that all the six models contain one or more linearly correlated parameters. Thus, statistically they suffer from the deflect due to collinearity. However, such a problem was thoroughly investigated by Randic [8]. We have, therefore, used Randic recommendations to resolve the problem arising from co- linearity. Randic [8] stated that selection of the descriptors to be used in structure-property-activity studies should not be delegated solely to the computers although statistical criteria will continue to be useful for

preliminary screening of the descriptors taken from a large pool. Often in an automated selection of descriptors a descriptor will be discarded because it is highly correlated with another descriptor already selected. But what is important is not descriptor parallel to one another, that is, duplicate much of the same structural information but whether they in those parts that are important for structure-property-activity correlation. If they differ in the domain, which is important for the property / activity considered both descriptors should be retained. If they differ in parts that are not relevant for the correlation of considered in parts that are not relevant for the correlation of considered property / activity that one of them can be discarded. Therefore, following Randic [8] all the six models can be considered statistically significant. In this regard it is worthy to mention that some of the most obvious problems of severe multicollinearity are as follows:

(1) Incorrect size of the coefficients,

(2) A change in the values of the previous coefficient when a new variable is added to the model,

(3) Change in insignificant of a preciously significant variable when a new variable is added to the model, and

(4) An increase in the standard error of the estimate when a new variable is added to the model.

In the proposed models (33-38) none of these problems occur. Furthermore, all the variables occurring in the model have coefficients, which are significantly larger than their respective standard deviations. In view of the aforementioned discussion all these models are considered statistically significant.

In order to finalize our results it is worthy to comment on the degeneracy of the molecular described in the present study. A perusal of Table 3 shows that low to high degeneracy is present in all the molecular descriptors used. This due to the fact that these descriptors belong to first and second generation descriptors , which in spite of their degeneracy are quite useful in QSPR and QSAR studies [15].In our case the degeneracy problem has become more actuate due to the use of descriptors LP1 and X5AV. Out of the 53 compounds used in the present study LP1 is found to be the same (2.639) value for as many as 49 compounds. While in the case of X5AV, its value is not widely varied, it ranges between 0.022 and 0.025.Furthermore both these parameters are involved in all the six (33-38) statistically significant models. However, we observed that use of these parameters in the proposed models is well justified due to increase in R^2_A upon their addition as the correlating parameters. Furthermore, in all the six models these parameters have coefficients very much larger than their corresponding standard divisions and those models are statistically allowed. However, it seems beneficial to confirm our results by further performing regressions without considering LP1 or X5AV or both. If, under such a study the quality of the regression is improved, then it will be better to do modeling without these parameters, otherwise not. When we did so (Tables 14-19) we observed that under such study the models become quite inferior without the use of these parameters. All these results, therefore, justifies the use of these two parameters in all the models proposed by us.

Modified models 33 – 38 after deleting LP1 or X5AV or both variables

Table 14. Regression parameters for models 33 – 38 after deleting LP1

Models	R	R^2	R^2_A	Standared Error (SE)	F	Q= R/SE
33	0.6929	0.6929	0.3570	0.5541	3.8998	1.2505
34	0.7105	0.5049	0.3710	0.5480	3.7727	1.2965
35	0.7267	0.5281	0.4005	0.5350	4.1402	1.3582
36	0.7270	0.5285	0.3844	0.5422	3.6679	1.3408
37	0.7038	0.4953	0.3589	0.5533	3.6314	1.2720
38	0.7840	0.6146	0.4969	0.4902	5.2195	1.5994

Table 15. Cross Validation parameters for models 33 – 38 after deleting LP1

Models	PRESS	S_{PRESS}	SSY	R^2_{CV}	PRESS/SSY	PSE
33	29.2527	0.8774	47.0000	0.3776	0.6224	0.7807
34	28.1088	0.8716	47.0000	0.4019	0.5981	0.7652
35	27.0600	0.8552	47.0000	0.4243	0.5757	0.7508
36	27.0424	0.8667	47.0000	0.4246	0.5754	0.7506
37	28.5472	0.8784	47.0000	0.3926	0.6074	0.7712
38	23.3385	0.8052	47.0001	0.5034	0.4966	0.6973

Table 16. Regression parameters for models 33 – 38 after deleting X5AV

Models	R	R^2	R^2_A	Standared Error (SE)	F	Q= R/SE
33	0.7207	0.5194	0.4056	0.5328	4.5638	1.3528
34	0.7397	0.5471	0.4247	0.5241	4.4701	1.4113
35	0.7405	0.5484	0.4263	0.5234	4.4930	1.4149
36	0.7648	0.5849	0.4581	0.5087	4.6118	1.5034
37	0.7629	0.5820	0.4690	0.5035	5.1516	1.5150
38	0.7829	0.6130	0.4947	0.4912	5.1833	1.5939

Table 17. Cross Validation parameters for models 33 – 38 after deleting X5AV

Models	PRESS	S_{PRESS}	SSY	R^2_{CV}	PRESS/SSY	PSE
33	27.4474	0.8499	47.0000	0.4160	0.5840	0.7562
34	26.2159	0.8417	47.0000	0.4422	0.5578	0.7390
35	26.1604	0.8409	47.0000	0.4434	0.5566	0.7382
36	24.5844	0.8264	47.0001	0.4769	0.5231	0.7157
37	24.7086	0.8172	47.0000	0.4743	0.5257	0.7175
38	23.4070	0.8063	47.0000	0.5020	0.4980	0.6983

Table 18. Regression parameters for models 33 – 38 after deleting LP1 and X5AV

Models	R	R^2	R^2_A	Standared Error (SE)	F	Q= R/SE
33	0.6574	0.4321	0.3156	0.5717	3.7096	1.1499
34	0.6877	0.4730	0.3482	0.5579	3.7893	1.2327
35	0.7028	0.4939	0.3740	0.5467	4.1200	1.2854
36	0.7030	0.4942	0.3575	0.5539	3.6151	1.2692
37	0.6866	0.4714	0.3462	0.5588	3.7652	1.2288
38	0.7387	0.5457	0.4229	0.5250	4.4437	1.4071

Table 19. Cross Validation parameters for models 33 – 38 after deleting LP1 and X5AV

Models	PRESS	S_{PRESS}	SSY	R^2_{CV}	PRESS/SSY	PSE
33	31.5632	0.8996	47.0000	0.3284	0.6716	0.8109
34	29.5900	0.8824	47.0000	0.3704	0.6296	0.7851
35	28.6141	0.8678	46.9999	0.3912	0.6088	0.7721
36	28.5991	0.8792	47.0000	0.3915	0.6085	0.7719
37	29.6652	0.8836	47.0001	0.3688	0.6312	0.7861
38	26.2803	0.8428	47.0001	0.4408	0.5592	0.7399

Conclusions

From the results and discussion made above we conclude that antimycobacterial activity of 3-formyl rifamycin SV derivatives can be modeled using a twelve-parametric model which contains variety of molecular descriptors including distance-based and connectivity indices. The results obtained here in will be useful for pharmaceutical as well as medicinal chemists to synthesis new drugs having still better antibacterial potential

Experimental Section

(1) Antimycobacterial activity: The antimycobacterial activity expressed as $\log(MIC^{RIA}/MIC^{X})$ for different strains against *Mycobacterium tuberculosis* were taken from the literature [4]. For the brevity this activity in all the tables as well as in the text is expressed as logA. Further details are available in [4].

(2) Molecular descriptors: All the molecular descriptors used for proposing statistically significant models were calculated using DRAGON Software [11] . The structure optimization was performed using ACD Labs [12] and HyperChem [13] software's.

(3) Statistical analysis: All the statistical analyses were performed using SPSS Software [14].

References and Notes

1. Zhang, Y.; Garbc, T.; Young, D., *Mol. Microbiol.* **1993**, *8*, 521.
2. Zhang, Y.; Heym, B.; Allen ,R.;Young, D., Cole, S. *Nature* **1992**, *358*, 591 .
3. Hartman, G.R.; Heinrich, P.; Kollenda, M.C.; Skrobranek, G.; Tropschung, M.; Weiff, W. *Angew. Chem.*, **1985**, *97*, 1011.
4. Dimov, D.; Nedyalkova, Z.; Haladjeva, S.; Schuticmann, G.; Mekenyan, O., QSAR Modeling of Antimicrobacterial Activity and Activity Against Other Bacteria of 3-Formyl Rifamycin SV Derivatives, *Quant. Struct. Act. Relat.* **2001**, *20*, 302.
5. Chatterjee, S.; Hadi, A.S.; Bice, B., Regression Analysis by Examples, John Wiley, New York (3rd Edn.), **2000**.
6. Box, G.E.P.; Hunter W.G.; Hunter, J.S., Statistics for Experimenters, John Wiley, New York, **1978**.
7. Diudea, M.V.; Florescu, M.S.; Khadikar, P.V., Molecular Topology and its Applications, Eficon Press, Bucuresti, Romania, **2006,** Chapter VI, pp. 169-216.
8. Randic, M., *Croat Chem. Acta*, **1993**, *66*, 289-295; *Acta Chem. Slov.* **1998** *45*, 239.
9. Tute, M.S. in Advances in Drug Research, Harper, N.J.; Simmond, A.B. (Eds), Academic Press, London, **1971**.
10. Crown, H., Comprehensive Medicinal Chemistry, Vol.4, Pergmon Press, New York, **1990**, p.19.
11. DRAGON Software for calculation of molecular descriptors, www.disat.unimib.it.
12. ACD Labs Software for structure optimization, Chem Sketch 3.0, www.acdlabs.com.
13. HyperChem6 Software for structure optimization, http://www.hyper.com
14. SPSS Software (version 13, SPSS, Inc.), http://www.spss.com
15. Balaban, A.T. *J.Chem. Inf. Comput. Sci.* **1992**, *32*, 23.