Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry 13 (2005) 343–351
Bioorganic & Medicinal Chemistry

CP-MLR directed QSAR studies on the antimycobacterial activity of functionalized alkenols—topological descriptors in modeling the activity

Manish K. Gupta, Ram Sagar, Arun K. Shaw and Yenamandra S. Prabhakar*

Division of Medicinal and Process Chemistry, Central Drug Research Institute, Lucknow 226 001(UP), India

Received 18 May 2004; revised 18 August 2004; accepted 9 October 2004

Abstract—The antimycobacterial activity of nitro/acetamido alkenol derivatives and chloro/amino alkenol derivatives has been analyzed through combinatorial protocol in multiple linear regression (CP-MLR) using different topological descriptors obtained from Dragon software. Among the topological descriptor classes considered in the study, the activity is correlated with simple topological descriptors (**TOPO**) and more complex 2D autocorrelation descriptors (**2DAUTO**). In model building the descriptors from other classes, that is, empirical, constitutional, molecular walk counts, modified Burden eigenvalues, and Galvez topological charge indices have made secondary contribution in association with **TOPO** and/or **2DAUTO** classes. The structure–activity correlations obtained with the **TOPO** descriptors suggest that less branched and saturated structural templates would be better for the activity. For both the series of compounds, in **2DAUTO** the activity has been correlated to the descriptors having mass, volume and/or polarizability as weighting component. In these two series of compounds, however, the regression coefficients of the descriptors have opposite arithmetic signs with respect to one another. Outwardly these two series of compounds appear very similar. But in terms of activity they belong to different segments of descriptor–activity profiles. This difference in the activity of these two series of compounds may be mainly due to the spacing difference between the C1 (also C6) substituents and rest of the functional groups in them.

1. Introduction

The mycobacterium, *Mycobacterium tuberculosis*, is the causative agent of tuberculosis (TB) and is responsible for the morbidity and mortality of significant population in the tropical and subtropical countries and its control is a high-priority task for both the developed as well as developing economies.[1] Traditionally, the clinical management of TB has relied heavily on a limited number of drugs such as isonicotinicacid hydrazide, rifampicin, ethambutal, streptomycin, ethionamide, pyrazinamide, fluoroquinolones, etc.[2,3] However, the prolonged treatment schedules and development of drug resistance strains of the mycobacterium complicated the clinical picture and prompted for the search of new chemical prototypes for its control.[4–6] In the development of chemotherapeutic strategies for the control of pathogenic agents, their outer cell walls and membranes serve as attractive targets for exploring new drugs—the mechanism of action of penicillin is a well-understood case in this direction.[7] In mycobacteria, the cell wall structure consists of a dense network of cross-linked sugar residues esterified with mycolic acid at the ends.[8,9] This understanding has prompted for the identification and development of new chemical prototypes with distinctly different characteristics than the existing ones with an aim to overcome their limitations.[10–14] Here, we have considered sugar templates as a potential source of new molecular scaffolds with strategically positioned functional groups that will selectively interact and communicate with the complementary groups/sites of the pathogen's cell-wall structure and receptor(s) therein. In this background, as a part of our laboratory's ongoing new drug development program on the design of antimycobacterial agents, some highly functionalized heptenol and octenol derivatives (Fig. 1) have been prepared by opening the glucal/galactal derivatives at the anomeric center and evaluated them for antimycobacterial activity.[15,16] To the best of our knowledge, these compounds were first of their class to show this kind of activity.[15,16] Here, establishing a correlation between

Keywords: Functionalized heptenol/octenol; Antimycobacterial activity; Combinatorial protocol in multiple linear regression (CP-MLR); Topological descriptors; QSAR study.
* Corresponding author. Tel.: +91 522 2212411; fax: +91 522 2223405; e-mail: yenpra@yahoo.com

0968-0896/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2004.10.025

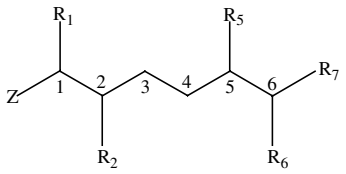

Figure 1. General structure of the functionalized alkenols associated with antimycobacterial activity. These compounds were derived starting from glucal or galactal derivatives by opening the respective pyranose rings at the anomeric carbon (Cl). For the purpose of correspondence with the parent sugars, in this structural template the carbon corresponding to the anomeric carbon of pyranose has been opted as '1' (that is Cl) and the succeeding carbons as '2' (C2), '3' (C3) etc. Here, for the structural template corresponding to nitro/acetamido alkenol derivatives (Table 2), R_5 and R_6 is OAc, and R_7 is CH2OAc and for that of chloro/amino alkenol derivatives (Table 3), R_5 is OBn and C3–C4 centers are joined with double bond.

the structure and the associated activity will come to aid in furthering the understanding of the system under investigation. In these studies, parameterization of chemical structure plays a pivotal role. In enumeration of chemical structures, it is important to note that in isolation a data point is only a qualified number. A collection of such qualified numbers makes a variable or descriptor. In mathematical models, each and every variable communicate with all other variables. A meaningful inter- and intra-variable communications result in the evolution of models with predictive value. With this philosophy, a quantitative structure–activity relationship (QSAR) study has been contemplated on functionalized heptenol and octenol derivatives (hereafter referred to as functionalized alkenol derivatives) (Fig. 1) reported earlier from this laboratory.[15,16] For this different empirical, constitutional, and topological descriptors of the compounds have been opted from Dragon software.[17] The descriptor classes considered in the study along with their definitions and scopes in addressing the molecular structure have been presented in Table 1. With all these descriptors as independent parameters, the structure–activity relations of the antimycobacterial activity of the functionalized alkenol derivatives (Tables 2 and 3) have been discovered using the combinatorial protocol in multiple linear regression (CP-MLR) approach, which is a variable selection procedure for model development in QSAR and QSPR studies.[18–20] As the total number of descriptors involved in this study are about 290 for each set of compounds, only the names of descriptor classes and participating descriptor names have been addressed in the discussion.[21] The details of computational procedure is discussed in the methodology. The results are presented here.

2. Results and discussion

In multi-descriptor class environment, exploring for models along the descriptor class provides an opportunity to understand the phenomenon under investigation vis-à-vis the concept embedded in the descriptor. With this, attempts have been made to develop one- and two-descriptor models for the antimycobacterial activity of the nitro/acetamido alkenol derivatives (Table 2) and chloro/amino alkenol derivatives (Table 3). For nitro/acetamido alkenol derivatives (Table 2), the study has led to four models in the topological (**TOPO**) class descriptors and 23 models in the 2D autocorrelations (**2DAUTO**) class descriptors (Table 1). In case of chloro/amino alkenol derivatives (Table 3), only the

Table 1. The descriptor classes used for the analysis of antimycobacterial activity of nitro/acetamido alkenol (Table 2) and chloro/amino alkenol (Table 3) derivatives and number of models identified in each class

Descriptor class (acronym)[a]	Definition and scope (number of descriptors)	Descriptors per model (number of models)[b]	
		Table 2	Table 3
Empirical (**EMP**)	Descriptors represent the counts of non-single bonds, hydrophilic groups and ratio of the number of aromatic bonds and total bonds in H-depleted molecule (2)		
Constitutional (**CONST**)	Dimensionless descriptors, independent from molecular connectivity and conformations (24)		
Topological (**TOPO**)	Descriptors obtained from molecular graphs and independent of conformations (68)	1(1); 2(3)	
Molecular walk counts (**MWC**)	Descriptors representing self-returning walk counts of different lengths (13)		
Modified Burden eigenvalues (**BCUT**)	Descriptors representing positive and negative eigenvalues of the adjacency matrix, weights the diagonal elements & atoms (64)		
Galvez topological charge indices (**GVZ**)	Descriptors representing the first 10 eigenvalues of corrected adjacency matrix (21)		
2D autocorrelations (**2DAUTO**)	Molecular descriptors calculated from the molecular graphs by summing the products of atom weights of the terminal atoms of all the paths of the considered path length (the lag) (96)	1(3); 2(20)	2(5)

[a] Ref. 17.

[b] Models emerged from CP-MLR protocol with filter-1 as 0.3; filter-2 as 2.0; filter-3 as 0.74; filter-4 as $0.3 \leqslant Q^2 \leqslant 1.0$, number of compounds in each dataset is 11.

Table 2. Antimycobacterial activity of nitro/acetamido alkenol derivatives (Fig. 1; for compounds **8** to **15** Z is acetamido group and for the remaining it is nitro group)

Compd[a] No.	R_1	R_2	C_5^*	$B_{3,4}$[b]	−log MIC			
					Obsd[c]	Eq. 1	Eq. 3	Eq. 8
1	H	OH	*S*	db	na	4.15	3.91	4.07
2	H	OH	*R*	db	4.12	4.15	3.91	4.07
3	CH_3	OH	*S*	db	3.84	3.79	3.80	3.74
4	H	H	*S*	db	na	3.82	3.67	3.68
5	H	H	*R*	db	3.50	3.82	3.67	3.68
6	CH_3	H	*S*	db	na	3.42	3.56	3.43
7	CH_3	H	*R*	db	3.52	3.42	3.56	3.43
8	H	OAc	*S*	sb	na	3.71	4.68	2.54
9	H	OAc	*R*	sb	na	3.71	4.68	2.54
10	CH_3	OAc	*S*	sb	na	3.37	4.57	1.53
11	CH_3	OAc	*R*	sb	na	3.37	4.57	1.53
12	H	H	*S*	sb	4.72	4.73	4.80	4.70
13	H	H	*R*	sb	4.72	4.73	4.80	4.70
14	CH_3	H	*S*	sb	4.44	4.43	4.29	4.33
15	CH_3	H	*R*	sb	4.14	4.43	4.29	4.33
16	H	H	*S*	db	4.70	4.47	4.48	4.66
17	H	H	*R*	db	na	4.47	4.48	4.66
18	CH_3	H	*S*	db	4.72	4.14	4.32	4.31
19	CH_3	H	*R*	db	3.82	4.14	4.32	4.31

[a] In compounds **4** to **7**, C1 and C2 are connected with double bond.
[b] Bond order between C3 and C4, 'sb' for single bond and 'db' for double bond.
[c] Ref. 15.

Table 3. Antimycobacterial activity of chloro/amino alkenol derivatives (Fig. 1)

Compd No.	Z	R_1	R_2	R_6	R_7	C_5^*	$B_{1,2}$[a]	−log MIC	
								Obsd[b]	Eq. 5
20	CH_2Cl	Ac	H	OBn	CH_2OBn	*S*	db	na	4.84
21	CH_2Cl	Ac	H	OAc	CH_2OBn	*R*	db	5.16	4.94
22	CH_2Cl	Ac	H	OAc	H	*S*	db	4.13	4.32
23	$CH_2N(CH_2)_5$[c]	Ac	H	OAc	CH_2OBn	*R*	db	4.00	4.35
24	$CH_2N(CH_2)_4$[d]	Ac	H	OAc	CH_2OBn	*R*	db	4.29	4.48
25	$CH_2N(C_2H_5)_2$	Ac	H	OAc	CH_2OBn	*R*	db	3.99	3.80
26	$CH_2N(C_2H_4)_2NMe$[e]	Ac	H	OAc	CH_2OBn	*R*	db	4.32	4.33
27	$CH_2N(CH_2)_5$[c]	Ac	H	OAc	H	*S*	db	3.89	3.85
28	$=CH_2$	CN	OH	OBn	CH_2OBn	*S*	sb	3.97	4.10
29	$=CH_2$	CN	OH	OH	CH_2OBn	*R*	sb	3.88	3.86
30	$CH_2N(C_2H_4)_2NMe$[e]	CN	OH	OBn	CH_2OBn	*S*	sb	4.66	4.35
31	$CH_2N(C_2H_4)_2NMe$[e]	CN	OH	OH	CH_2OBn	*R*	sb	4.28	4.20

[a] Bond order between C1 and C2, 'sb' for single bond and 'db' for double bond.
[b] Ref. 16.
[c] Piperidinyl methyl.
[d] Pyrrolidinyl methyl.
[e] N-methyl-piperazinyl methyl.

2DAUTO class descriptors have resulted in five models (Table 1). The essence of all the models obtained from the study has been provided in Tables 4 and 5 in the form of identified descriptor's average of regression coefficient along with its standard deviation across the models and the total incidence corresponding to all the models. This, while providing the averages of the estimated regression coefficients of all the identified descriptors, shows their variance across the models emerged from the study as well. To maintain brevity, the complete regression equations have been shown for selected models only. The following regression equation represents a **TOPO** class structure–activity model of the nitro/acetamido alkenol derivatives (Table 2).

$$-\log \mathrm{MIC} = 9.691 - 0.034(0.016)\mathrm{TWC} - 0.824(0.192)\mathrm{PCR}$$
$$n = 11,\ r = 0.840,\ Q^2 = 0.487,\ s = 0.296,\ F = 9.57 \quad (1)$$

In this and all other regression equations, n is the number of compounds, r is the correlation coefficient, Q^2 is cross-validated R^2 from leave-one-out (LOO) procedure, s is the standard error of the estimate and F is the F-ratio between the variances of calculated and observed activities. The values given in the parentheses are the standard errors of the regression coefficients. Also, in the randomization study (100 simulations per

Table 4. Descriptors identified in modeling the antimycobacterial activity of nitro/acetamido alkenol derivatives (**Table 2**) along with their average regression coefficients

Descriptor[a]	Av Reg Coef (sd) Total incidence[b]		Descriptor[a]	Av Reg Coef (sd) Total incidence[b]	
	Individual	Combined		Individual	Combined
EMP			**2DAUTO**		
Ui		−1.918(0.170)2	ATS3m	−7.584(—)1	
Hy		−0.561(0.080)2	ATS1v	−46.804(—)1	
			ATS3v	−27.444(0.670)2	
CONST			ATS7v	−32.081(—)1	
Mv		−31.900(—)1	ATS8v	−31.892(8.200)2	
nDB		−0.479(0.040)2	ATS2p	−67.242(—)1	
nOH		−0.326(0.050)4	ATS3p	−36.461(0.582)2	
nHD		−0.294(0.026)4	ATS4p		−13.241(—)1
			ATS7p		−43.365(—)1
TOPO			ATS8p	−21.634(3.031)2	
TWC	−0.034(—)1		MATS1m	17.082(0.833)4	
TPCM	−0.003(—)1		MATS3m	6.863(1.361)2	
PCR	−0.802(0.0285)3		MATS6m	−6.116(0.929)4	
PCD		−0.244(0.016)2	MATS7m	−17.016(1.092)2	
SMTIV	−0.0004(—)1		MATS1v	−12.547(2.707)3	
GSI	−0.162(—)1		MATS4v		7.761(—)1
			MATS3e		7.743(—)1
MWC			MATS1p	−8.622(—)1	
MWC04		−0.195(—)1	GATS6m	3.206(—)1	
			GATS7m	5.959(0.502)3	
BCUT			GATS1v	3.536(0.007)2	
BEPm2		−11.317(0.015)2	GATS2v	4.697(—)1	
BEPv2		−6.728(1.346)4	GATS1e	−1.653(0.023)2	
BEPe2		−10.858(1.257)3	GATS2e	−4.245(—)1	
BEPp2		−6.225(0.099)2	GATS1p	3.131(0.007)2	
			GATS2p	4.577(0.604)2	
			GATS3p	−3.256(—)1	
GVZ					
GGI3		−1.357(—)1			
JGI3		−35.204(—)1			

[a] The descriptors are identified from the one and two parameter models emerged from CP-MLR protocol with filter-1 as 0.3; filter-2 as 2.0; filter-3 as 0.74; filter-4 as $0.3 \leqslant Q^2 \leqslant 1.0$, number of compounds in the study are eleven; **EMP:** Ui is unsaturation index; Hy is hydrophilic factor; **CONST:** Mv is mean atomic van der Walls volume; nDB is number of double bonds; nOH is number of hydroxyl groups; nHD is number of donor atoms for H bond; *n*Cl is number of chloro groups present in the molecule; **TOPO:** TWC is total walk count; TPCM is total multiple walk count; PCR is ratio of multiple path counts to path counts; PCD is difference of multiple path counts to path counts; SMTIV is Schultz molecular topological index by valence vertex degrees; GSI is Gordon–Scantlebury index; **MWC:** MWC04 is molecular walk count of order 4; **BCUT:** in BEPwn, 'BEP' represents Positive Burden Eigenvalue; penultimate character '*w*' indicates the weighting property used in the computation which may be either '*m*', that is, atomic masses or '*v*', that is, atomic van der Walls volumes or '*e*', that is, Sanderson electronegativities or '*p*', that is, atomic polarizabilities; the last character '*n*' indicates the eigenvalue level; **GVZ:** GGI3 is topological charge index of order 3; JGI3 is mean topological charge index of order 3; **2DAUTO:** In the descriptors of autocorrelation of topological structure of ATSnw (Broto-Moreau), MATSnw (Moran) and GATSnw (Geary) the penultimate character, '*n*' indicates the autocorrelation vector of lag *n* corresponding to the number of edges in the fragment unit considered in the computation and the last character '*w*' refers to the same as in case of **BCUT**. Also see Ref. 17.

[b] The average regression coefficient of the descriptor corresponding to all models, its standard deviation (s.d.) and the total number of its incidence. The arithmetic sign of the coefficient represents the actual sign of the regression coefficient in the models. A blank in 'Individual' column indicates that descriptor has not participated in any model. The 'Combined' column corresponds to the dataset of all descriptor classes together and shows only those descriptors which additional to the 'individual' column.

model) none of the identified models has shown any chance correlation. The four models, including the above presented one, emerged from **TOPO** class have shared five descriptor among themselves (Table 4). In Eq. 1, the activity is correlated with ratio of multiple path counts to path counts (PCR) alone to the order of 0.74. In addition to this, activity is correlated with one more **TOPO** descriptor, that is, total multiple walk count (TPCM; $r = 0.79$). Also, all the participating **TOPO** class descriptors' regression coefficients have been associated with negative sign (Table 4). Among the descriptors, PCR has the highest incidence of three (Table 4). The emerged models suggest that a smaller PCR value would be favorable for the activity, which in other words favors more saturated system for better activity. The negative regression coefficient of TPCM, which accounts for the multiple bonded centers in the system, also suggests that compounds with less multiple bonds are better for the activity. The regression coefficients of total walk count (TWC) and Gordon–Scantlebury index (GSI) suggest that a compound with branching will be less favorable when compared to its linear homologue. Here, Schultz molecular topological index by valence vertex degrees (SMTIV) is an index that addresses both branching and multiple bonded centers in a given compound. In homologous series of com-

Table 5. Descriptors identified in modeling the antimycobacterial activity of chloro/amino alkanol derivatives (Table 3) along with their average regression coefficients

Descriptor[a]	Av reg coef (sd) total incidence	
	Individual	Combined
CONST		
*n*Cl		0.702(—)1
BCUT		
BEPm1		26.200(—)1
BEPm7		2.144(0.000)2
2DAUTO		
ATS8m	6.538(—)1	
ATS3p	13.778(—)1	
ATS7p	11.090(—)1	
ATS3v	13.778(—)1	
ATS7v	11.002(—)1	
GATS4m	−3.381(0.508)5	

[a] See footnotes 'a' and 'b' of Table 4.

$$-\log \mathrm{MIC} = 10.780 - 19.491(9.572)\mathrm{ATS8p} - 16.243(4.004)\mathrm{MATS7m}$$
$$n = 11,\ r = 0.869,\ Q^2 = 0.604,\ s = 0.270,\ F = 12.31 \quad (2)$$

$$-\log \mathrm{MIC} = -6.492 - 5.324(2.589)\mathrm{MATS6m} + 3.541(0.797)\mathrm{GATS1v}$$
$$n = 11,\ r = 0.865,\ Q^2 = 0.614,\ s = 0.273,\ F = 11.92 \quad (3)$$

$$-\log \mathrm{MIC} = -5.855 - 5.326(2.605)\mathrm{MATS6m} + 3.136(0.712)\mathrm{GATS1p}$$
$$n = 11,\ r = 0.863,\ Q^2 = 0.603,\ s = 0.275,\ F = 11.72 \quad (4)$$

pounds, the multiple bonded centers markedly increase the value of SMTIV whereas branching effect decreases the same but relatively to a lesser extent. This in turn suggests that the unsaturation in the compounds is more detrimental to the activity when compared to the branching effect. All these descriptors collectively indicate a preference for compounds with more saturated and less branched structural templates for antimycobacterial activity.

The **2DAUTO** class descriptors also represent the topological structure of the compounds. But these descriptors are more complex in nature when compared to the **TOPO** class descriptors. The **2DAUTO** descriptors considered in the study have their origin in autocorrelation of topological structure of Broto-Moreau (ATS), of Moran (MATS) and of Geary (GATS).[17] The computation of these descriptors involve the summations of different autocorrelation functions corresponding to the different fragment lengths and lead to different autocorrelation vectors corresponding to the lengths of the structural fragments.[22] Also a weighting component in terms of a physicochemical property has been embedded in this descriptor. As a result these descriptors address the topology of the structure or parts thereof in association with a selected physicochemical property. In the descriptor's nomenclature, the penultimate character, a number, indicates the number of consecutively connected edges considered in its computation and is called as the autocorrelation vector of lag *n* (corresponding to the number of edges in the unit fragment). The very last character of the descriptor's nomenclature indicates the physicochemical property considered in the weighting component—m for mass or v for volume or e for Sanderson electronegativity or p for polarizability—for its computation. The philosophy embedded in these descriptors and their computational aspects are available in different sources.[17,22] The 23 models emerged from this descriptor class have shared 23 descriptors among themselves (Table 4). The following three regression equations (Eqs. 2–4) represent the selected models of this class of descriptors.

In all the three models, the activity is correlated to a significant extent ($r = 0.73$ to 0.79) with one descriptor, which is either Moran autocorrelation lag seven weighed by atomic masses (MATS7m) or Geary autocorrelation lag one weighed by van der Walls volumes (GATS1v) or Geary autocorrelation lag one weighed by atomic polarizabilities (GATS1p). PCR, TPCM of **TOPO** class and these three descriptors are highly intercorrelated with each other ($r = 0.90$ to 0.99). Furthermore, the trend of the formed models and descriptors participated therein (Table 4) indicate that descriptors of lag one have accounted for the maximum number of models (cumulative frequency is fifteen) followed by lag three (cumulative frequency is eight) and lag seven (cumulative frequency is six) descriptors. A study of the correlation matrix of these descriptors revealed that the autocorrelation vectors of each lag formed by mass, volume, and polarizability weightings are highly intercorrelated. Also, in addition to the three descriptors (MATS7m, GATS1v, and GATS1p) of Eqs. 2–4, seven more descriptors (ATS3v, ATS7v, ATS2p, ATS3p, MATS1p, and GATS7m) have shown correlation individually with the activity to the order of 0.70–0.79. In this class, the lag one descriptors are simplest among all. They do not differentiate between normal and branched structural templates, but identify the loops and multi-pully connected vertices in them. These descriptors also suggest that unsaturation in the structural templates is not favorable for the activity. The participation of descriptors of lag six and seven may be viewed in terms of association of activity information content with the six and seven centered structural fragments. Most of the identified descriptors have mass, volume and/or polarizability as physicochemical weighting component in them indicating their influence on the activity. However, further deciphering of the information content of these descriptors is very complex as their computations involve integration of the structural fragments and due to this it is not possible to traverse backward from a higher state to a lower one.[22] In Eqs. 1–4 and for all other identified models the r and Q^2 values are not of very high order. It is mainly due to

the deviations in the calculated as well as predicted activities of compounds **18** and **19**. Exclusion of these two compounds from the dataset improved the Eqs. 1–4 as shown in Eqs. 1a,2a,3a,4a, respectively.

$$-\log \mathrm{MIC} = 9.797 - 0.035(0.011)\mathrm{TWC} - 0.834(0.132)\mathrm{PCR}$$
$$n = 9,\ r = 0.935,\ Q^2 = 0.638,\ s = 0.203,\ F = 20.725 \quad (1a)$$

$$-\log \mathrm{MIC} = 11.281 - 21.071(6.443)\mathrm{ATS8p} - 15.806(2.566)\mathrm{MATS7m}$$
$$n = 9,\ r = 0.957,\ Q^2 = 0.767,\ s = 0.166,\ F = 32.58 \quad (2a)$$

$$-\log \mathrm{MIC} = -6.523 - 5.477(1.690)\mathrm{MATS6m} + 3.554(0.511)\mathrm{GATS1v}$$
$$n = 9,\ r = 0.952,\ Q^2 = 0.781,\ s = 0.175,\ F = 29.06 \quad (3a)$$

$$-\log \mathrm{MIC} = -5.858 - 5.307(1.752)\mathrm{MATS6m} + 3.136(0.469)\mathrm{GATS1p}$$
$$n = 9,\ r = 0.948,\ Q^2 = 0.763,\ s = 0.181,\ F = 26.83 \quad (4a)$$

However, except the deviations in the calculated/predicted activities we have no other reason to exclude compounds **18** and **19** from the analysis. The regression coefficients of all the descriptors in Eqs. 1a–4a are in general agreement with those of the Eqs. 1–4 as well as with those presented in Table 4.

The five models of chloro/amino alkenol derivatives (Table 3) have shared six **2DAUTO** class descriptors among themselves (Table 5). The following equations (Eqs. 5 and 6) represent the selected structure–activity models of these compounds.

$$-\log \mathrm{MIC} = 7.068 + 11.002(3.571)\mathrm{ATS7v} - 3.509(0.892)\mathrm{GATS4m}$$
$$n = 11,\ r = 0.856,\ Q^2 = 0.443,\ s = 0.224,\ F = 10.95 \quad (5)$$

$$-\log \mathrm{MIC} = 6.996 + 6.538(2.243)\mathrm{ATS8m} - 2.512(0.931)\mathrm{GATS4m}$$
$$n = 11,\ r = 0.846,\ Q^2 = 0.418,\ s = 0.230,\ F = 10.09 \quad (6)$$

For all the models emerged here, Geary autocorrelation lag four weighed by atomic masses (GATS4m) is one common descriptor. The remaining five descriptors have come from Broto-Moreau autocorrelation of topological structure (ATS) with mass (m) or volume (v) or polarizability (p) as the weighting parameter (Table 5). All these five ATS descriptor are highly intercorrelated ($r = 0.86$ to 0.99). The correlation of activity with autocorrelation vectors of lags three, four, and seven indicate that structural fragment of corresponding lengths are enriched with activity information. At the same time, these descriptors indicate the role of physicochemical properties such as mass, volume, and/or polarizability of the compounds in deciding the activity. In this dataset, exclusion of compound **23** further improved the quality of the models as shown by Eqs. 5a and 6a, which correspond to Eqs. 5 and 6, respectively.

$$-\log \mathrm{MIC} = 7.482 + 11.098(3.112)\mathrm{ATS7v} - 3.716(0.785)\mathrm{GATS4m}$$
$$n = 10,\ r = 0.903,\ Q^2 = 0.588,\ s = 0.195,\ F = 15.41 \quad (5a)$$

$$-\log \mathrm{MIC} = 7.510 + 6.179(2.212)\mathrm{ATS8m} - 2.693(0.922)\mathrm{GATS4m}$$
$$n = 10,\ r = 0.868,\ Q^2 = 0.470,\ s = 0.225,\ F = 10.70 \quad (6a)$$

As the order and magnitude of the regression coefficients have remained the same, the equations formed by exclusion of compound(s) also convey the same information as their predecessors.

The structure–activity relations of the antimycobacterial activity of the nitro/acetamido alkenol derivatives (Table 2) and chloro/amino alkenol derivatives (Table 3) have been further investigated by creating two more datasets (datasets 'A' and 'B') for each series of compounds. Of these, dataset 'A' corresponds to all descriptor classes, which have not yielded any model individually. This has been chosen to investigate the collective information content of left out descriptor classes vis-à-vis the activity of the compounds. And the other, dataset 'B', contains all the descriptor classes that is selected as well as left out classes of the study. Under the study conditions, the dataset 'A' of nitro/acetamido alkenol derivatives (Table 2) did not result in any model. However, the dataset 'A' corresponding to chloro/amino alkenol derivatives (Table 3) has resulted in only one two-descriptor model (Eq. 7). Equation 7a is a variant of Eq. 7 formed by excluding compound **23** from the dataset.

$$-\log \mathrm{MIC} = -2.439 + 0.702(0.182)n\mathrm{Cl} + 2.144(0.578)\mathrm{BEPm7}$$
$$n = 11,\ r = 0.857,\ Q^2 = 0.318,\ s = 0.223,\ F = 11.03 \quad (7)$$

$$-\log \mathrm{MIC} = -2.962 + 0.679(0.167)n\mathrm{Cl} + 2.327(0.540)\mathrm{BEPm7}$$
$$n = 10,\ r = 0.893,\ Q^2 = 0.480,\ s = 0.203,\ F = 13.84 \quad (7a)$$

This clearly indicates that the descriptor classes other than **TOPO** and/or **2DAUTO** classes have insufficient or marginal information (corresponding to the activity) within themselves—independently as well as collectively—to come up to the level of model formation. However, under the same conditions, dataset 'B' of nitro/acetamido alkenol derivatives (Table 2) and chloro/amino alkenol derivatives (Table 3) have resulted in 61 and 7 models, respectively. In dataset 'B', the models obtained over and above the **TOPO** and/or **2DAUTO** individual classes indicate the cooperative participation and contribution of dataset 'A' in combination with the **TOPO** and/or **2DAUTO** classes. All the descriptors of the additional models emerged from the dataset 'B' of the compounds of Tables 2 and 3 have been listed in Tables 4 and 5, respectively. In nitro/acetamido alkenol derivatives (Table 2), the empirical (**EMP**) (two descriptors), constitutional (**CONST**) (four descriptors), molecular walk counts (**MWC**) (one descriptor), modified Burden eigenvalues (**BCUT**) (four descriptors) and Galvez topological charge indices (**GVZ**) (two descriptors) have participated in the models formation in combination with **TOPO** and **2DAUTO** classes. In this process, one more descriptor of **TOPO** and four more descriptors of **2DAUTO** class have joined the descriptor lists of respective classes as contributing ones. With respect to activity, these five descriptors may be carrying marginal information. Interestingly, the participating empirical and constitutional descriptors also favor more saturated and hydrophobic structural scaffolds for better activity (Table 4). All the four participating **BCUT** class descriptors correspond only to positive Burden eigenvalue-2. Among these, BEPm2 is correlated with the activity to the order of 0.74. All the four descriptors are intercorrelated ($r \sim 0.60$ to 0.99). Many of the descriptors listed in Table 4 have been associated with negative regression coefficients. It suggests that activity profile of nitro/acetamido alkenol derivatives (Table 2) and the identified descriptors have inverse proportionality relationship. If one views the compounds in terms of a normal distribution curve of their activity with respect to the molecular descriptors, all these analogues correspond to the post-optimum region of the said curve. Equation 8 with **BCUT** and **2DAUTO** class descriptors has emerged as the best model of dataset 'B' of nitro/acetamido alkenol derivatives (Table 2). This has improved further on the exclusion of compounds **18** and **19** (Eq. 8a).

$$-\log \mathrm{MIC} = 25.609 - 6.278(2.627)\mathrm{BEPv2} - 18.902(3.729)\mathrm{MATS7m}$$
$$n = 11,\ r = 0.884,\ Q^2 = 0.678,\ s = 0.254,\ F = 14.42 \quad (8)$$

$$-\log \mathrm{MIC} = 25.369 - 6.205(1.384)\mathrm{BEPv2} - 19.045(1.978)\mathrm{MATS7m}$$
$$n = 9,\ r = 0.973,\ Q^2 = 0.864,\ s = 0.133,\ F = 52.64 \quad (8a)$$

For chloro/amino alkenol derivatives (Table 3), the dataset 'B' has resulted in no better model than those from **2DAUTO** class alone. However, due to the synergetic contributions of a constitutional (**CONST**) and two modified Burden eigenvalue (**BCUT**) descriptors, two additional models have emerged from this dataset and Eq. 7 is one of them. The participating **CONST** descriptor suggests that chloro substitution on the structural template leads to better activity (Table 5).

3. Conclusion

In summary, the antimycobacterial activity of both nitro/acetamido alkenol derivatives (Table 2) and chloro/amino alkenol derivatives (Table 3) have been found to be correlated to simple (**TOPO**) to complex (**2DAUTO**) topological descriptors of these compounds. The descriptors from other classes, that is, empirical, constitutional and remaining topological have made secondary contribution in association with **TOPO** and/or **2DAUTO** classes. Exclusion of compounds **18** and **19** from the dataset of nitro/acetamido alkenol derivatives (Table 2) and compound **23** from that of chloro/amino alkenol derivatives (Table 3) has improved the significance of all identified models. However, these exclusions did not skew the order and magnitude of the regression coefficients of the descriptors. This lends credence to the models and descriptors identified therein. The antimycobacterial activities predicted by the selected models of this study are in agreement with the observed ones (Tables 2 and 3). In case of nitro/acetamido alkenol derivatives (Table 2), the correlations obtained with the **TOPO** descriptors suggest that less branched and saturated structural templates would be better for the activity. For both the series of compounds, in **2DAUTO** descriptors, the activity has been correlated to the autocorrelation vectors with mass, volume, and/or polarizability as weighting component. While most of the models of nitro/acetamido alkenol derivatives (Table 2) involved descriptors of lag one, six, and seven, for chloro/amino alkenol derivatives (Table 3) this happened to be from lag four. Four descriptors (ATS3p, ATS7p, ATS3v, and ATS7v) out of the six of **2DAUTO** class of chloro/amino alkenol derivatives (Tables 3 and 5) also appear in the descriptor list of nitro/acetamido alkenol derivatives (Tables 2 and 4). In these two series of compounds, however, the regression coefficients of these descriptors have opposite arithmetic signs with respect to one another. Also, unlike the nitro/acetamido alkenol derivatives (Table 2), for chloro/amino alkenol derivatives (Table 3) most of the contributing descriptors have been associated with positive regression coefficients. Outwardly these two series of compounds appear very similar. But in terms of activity they belong to different segments of descriptor–activity profiles. This difference in the activity of these two series of compounds may be mainly due to the spacing difference between the C1 (also C6) substituents and rest of the functional groups in them (Figure 1). This understanding will be helpful in modulating the spacing between the functional groups. The descriptors identified in the study contribute to the understanding of the designing of new/novel compounds in terms of optimum

molecular fragment lengths and physicochemical properties to be associated with them.

4. Methods

4.1. Dataset

Two series of compounds, nitro/acetamido alkenol derivatives (Table 2)[15] and chloro/amino alkenol derivatives (Table 3)[16] have been taken up in this study along with their antimycobacterial activity as logarithm of the inverse of minimum inhibitory concentration ($-\log$MIC where MIC in moles per liter against *M. tuberculosis*, $H_{37}R_v$). As these compounds (Fig. 1) have been synthesized starting from either glucal or galactal derivatives by opening the respective pyranose ring at the anomeric carbon, their main backbone has at least six carbons in them. To maintain correspondence with the parent templates, the numbering of main carbon backbone of these compounds (Fig. 1) has been done according to the numbering convention of pyranose sugar moiety (anomeric carbon as C1, and so on). For the computation of molecular descriptors, the structures of the compounds have been generated using the following protocol. As the compounds **12**, **13**, **20**, and **21** (Tables 2 and 3) are some of less substituted ones among the compounds of the study, they have been considered as the core templates (core compounds) for rest of the analogues. The 2D ChemDraw structure of core template has been converted into a 3D object using the default conversion procedure implemented in the CS Chem3D Ultra.[23] The so generated 3D-conformer of the compound has been subjected to energy minimization in the MOPAC module, using the AM1 procedure for closed shell systems, implemented in the CS Chem3D Ultra.[23] To maintain a well-defined conformer relationship across the compounds of the study, the structures of all other compounds have been generated by appending appropriate changes to the core template followed by 3D conversion and energy minimization as implemented in case of core compounds. In this study, structural models of compounds **1**, **3**, **4**, **6**, **8**, **10**, **14**, **16**, **18** (Table 2), **2**, **5**, **7**, **9**, **11**, **15**, **17**, **19** (Table 2), **22**, **27**, **28**, **30** (Table 3) and **23**–**26**, **29**, **31** (Table 3) have emerged from the core template of compounds **12**, **13**, **20**, and **21**, respectively. In the structure building process of all the compounds, the disposition of all functional groups at the asymmetric centers (except the ones corresponding to C1 and C2) and at the double bounds, where applicable, have been maintained according to our synthetic report.[15,16] In the absence of any information on the preferences of configurations of C1 and C2, they have been arbitrarily kept as S and R, respectively. This has no qualitative or quantitative effect on the parameters of the dataset of the compounds, as these structures have been used only to generate 2D descriptors. All these Chem3D Ultra generated structures of the compounds have been converted into Tripos MOL2 format (.ML2 files) in HyperChem[24] to enable them to be portable to Dragon software[17] for computing the various 2D molecular descriptors. The CP-MLR computational procedure[19] and model validation are briefly described below.

4.2. Computational procedure and model development

CP-MLR is a 'filter' based variable selection procedure for model development in QSAR and QSPR studies.[18–20] In this procedure, a combinatorial strategy with appropriately placed 'filters' has been interfaced with MLR to result in selected subset regressions for the extraction of diverse structure–activity models, each having unique combination of descriptors from the dataset under study. The 'filters' embedded in CP-MLR are aimed at (a) containing inter-parameter correlation to a predefined cutoff value (filter-1) (default acceptable value $\leqslant 0.3$), (b) optimize the variables entry to a model through *t*-values of regression coefficients (filter-2) (default acceptable value $\geqslant 2.0$), (c) comparability of models (equations) with different number of variable in terms of square root of adjusted multiple correlation coefficient of regression equation (filter-3) (r-bar; default acceptable value $\geqslant 0.74$), and (d) addressing the external consistency of the model with leave-one-out (LOO) cross-validation as default option (filter-4) (cross-validated R^2 or Q^2 criteria; default acceptable limits are $0.3 \leqslant Q^2 \leqslant 1.0$). They make the variable selection process efficient and lead to unique solutions. Furthermore, to discover any chance correlations associated with the models identified in CP-MLR, each cross-validated model has been subjected to randomization test[20,25] by repeated randomization of the biological response. The datasets with randomized response vector have been reassessed by multiple regression analysis. The emerging regression equations, if any, with correlation coefficients better than or equal to the one corresponding to unscrambled response data were counted. Every model has been subjected to 100 such simulation runs. This has been used as a measure to express the percent chance correlation of the model under scrutiny.[20] The CP-MLR protocol has been applied with default filter thresholds on the two datasets to identify the all possible models that could emerge from the descriptors of each set of compounds.

Supplementary data

Supplementary data associated with this article will be provided on request in notepad files; CDRI Communication No. 6572. Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2004.10.025.

References and notes

1. Dye, C.; Scheele, S.; Dolin, P.; Pathania, V.; Raviglione, M. C. *JAMA* **1999**, *282*, 677.
2. Khasnobis, S.; Escuyer, V. E.; Chatterjee, D. *Expert Opin. Ther. Targets* **2000**, *6*, 21.
3. Somoskovi, A.; Parsons, L. M.; Salfinger, M. *Respir. Res.* **2001**, *2*, 164.
4. Teodori, E.; Dei, S.; Scapecchi, S.; Gualtieri, F. *Il Farmaco* **2002**, *57*, 385.

5. Frieden, T. R.; Sterling, T. R.; Munsiff, S. S.; Watt, C. J.; Dye, C. *Lancet* **2003**, *362*, 887.
6. Smith, C. V.; Sharma, V.; Sacchettini, J. C. *Tuberculosis* **2004**, *84*, 45.
7. Bayles, K. W. *Trends Microbiol.* **2000**, *8*, 274.
8. McNeil, M.; Wallner, S. J.; Hunter, S. W.; Brennan, P. J. *Carbohydr. Res.* **1987**, *166*, 299.
9. Daffe, M.; Brennan, P. J.; McNeil, M. *J. Biol. Chem.* **1990**, *265*, 6734.
10. Jones, P. B.; Parrish, N. M.; Houston, T. A.; Stapon, A.; Bansal, N. P.; Dick, J. D.; Townsend, C. A. *J. Med. Chem.* **2000**, *43*, 3304.
11. Pasqualoto, K. F. M.; Ferreira, E. I. *Curr. Drug Targets* **2001**, *2*, 427.
12. Pathak, A. K.; Pathak, V.; Maddry, J. A.; Suling, W. J.; Gurcha, S. S.; Besra, G. S.; Reynolds, R. C. *Bioorg. Med. Chem.* **2001**, *9*, 3145.
13. Savini, L.; Chiasserini, L.; Gaeta, A.; Pellerano, C. *Bioorg. Med. Chem.* **2002**, *10*, 2193.
14. Katz, A. H.; Caufield, C. E. *Curr. Pharm. Design* **2003**, *9*, 857.
15. Pathak, R.; Shaw, A. K.; Bhaduri, A. P.; Chandrasekhar, K. V. G.; Shrivastava, A.; Shrivastava, K. K.; Chaturvedi, V.; Shrivastava, R.; Shrivastava, B. S.; Arora, S.; Sinha, S. *Bioorg. Med. Chem.* **2002**, *10*, 1695.
16. Pathak, R.; Pant, C. S.; Shaw, A. K.; Bhaduri, A. P.; Gaikwad, A. N.; Sinha, S.; Shrivastava, A.; Shrivastava, K. K.; Chaturvedi, V.; Shrivastava, R.; Shrivastava, B. S. *Bioorg. Med. Chem.* **2002**, *10*, 3187.
17. Dragon software (version 1.11-2001) by Todeschini, R.; Consonni, V. Milano, Italy and references cited therein. http://disat.unimib.it/chm/Dragon.htm.
18. Prabhakar, Y. S. *QSAR Comb. Sci.* **2003**, *22*, 583.
19. Prabhakar, Y. S. *Internet Electron. J. Mol. Des.* **2004**, *3*, 150, http://www.biochempress.com.
20. Prabhakar, Y. S.; Solomon, V. R.; Rawal, R. K.; Gupta, M. K.; Katti, S. B. *QSAR Comb. Sci.* **2004**, *23*, 234.
21. The complete descriptor dataset of all compounds will be provided on request.
22. Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem.* **1984**, *19*, 66.
23. ChemDraw Ultra 6.0 and Chem3D Ultra, Cambridge Soft Corporation, Cambridge, USA.
24. HyperChem Release 3, **1993**, Autodesk, California, USA.
25. So, S. S.; Karplus, M. *J. Med. Chem.* **1997**, *40*, 4360.