Topological estimation of cytotoxic activity of some anti-HIV agents: HEPT analogues

VIJAY K AGRAWAL[1], KAMLESH MISHRA[1], RUCHI SHARMA[1] and P V KHADIKAR[2]*

[1]QSAR and Computer Chemical Laboratories, APS University, Rewa 486 003, India
[2]Research Division, Laxmi Fumigation and Pest Control Pvt. Ltd., 3, Khatipura, Indore 452 007, India
e-mail: vijay-agrawal@lycos.com; pvkhadikar@rediffmail.com

Abstract. QSAR studies on anti-HIV cytotoxic activities of a series of HEPT(1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)-thymine) analogues have been discussed. The molecular descriptors used being van der Waals volume (Vw) and equalized electronegativity (c_{eq}). The *in vitro* cytotoxicities (pCC_{50}) were modelled using these parameters. It was observed that upon introduction of indicator parameters statistically excellent models are obtained. The predictive power of the models was examined using a cross-validation method.

Keywords. Anti-HIV agents; molecular modelling; van der Waals volume; cytotoxic activity; HEPT analogues; equalized electronegativity.

1. Introduction

The use of topological indices in modelling the toxicity of organic compounds is well known.[1–4] Of course, for modelling toxicity of chemicals one may employ molecular descriptors other than topological indices[5–10]. In our earlier study[11–13] we observed that the equalized electronegativity (c_{eq}),[14,15] van der Waals volume (Vw),[16] their combinations with each other and introduction of indicator parameters resulted in statistically excellent models. Such non-topological modelling appears to be interesting.

Reverse transcriptase (RT) plays a central role in the replication of HIV. A number of RT-inhibitors active against both HIV-1 and HIV-2 RT or only against HIV-1 RT have been discussed in the literature.[1,2]

The case of topological indices in modelling toxicity of organic compounds is well known.[3,4] Of course, for modelling toxicity of chemicals one may also employ molecular descriptors other than topological indices.[5–10] In our earlier study[11–13] we observed that equalized electronegativity[14,15] (c_{eq}), van der Waals volume[16] (Vw), their combinations with each other and introduction of indicator parameter(s) gave statistically significant models. The non-topological molecular descriptors gave fairly interesting results.

In this paper we have, therefore, carried out quantitative structure-activity relationship (QSAR) analysis on a series of HEPT (1-[(2-hydroxyethoxy) methyl]-6-(phenylthio)-thymine) analogues for modelling their cytotoxic activity using van der Waals volume (Vw), equalized electronegativity (c_{eq}) along with some indicator parameters. The details are given below. A series of 48 HEPT analogues as demonstrated in table 1 (figure 1) were used for this purpose.

The *in vitro* cytotoxicities (pCC_{50}) needed for the study were adopted from the literature.[3] The molecular descriptors, viz. equalized electronegativity (c_{eq}), and van der Waals volume (Vw) were calculated using the procedure described in the subsequent sections.[4–10] In addition, we have also used three dummy parameters (indicator parameters) Ip_1, Ip_2 and Ip_3 related to substituents at R^1, R^2 and R^4. Thus, we have a set of six molecular descriptors for modelling anti-HIV activity of 48 compounds in the present study. Our objective in the present study to determine which out of these six descriptors is useful in modelling the activity (pCC_{50}).

2. Methodology used

2.1 *Toxicity data*

In vitro cytotoxicities (pCC_{50}) were adopted from the literature.[17]

Table 1. Structural details and cytotoxic activity (pCC_{50}) of the compounds (HEPT analogues) used in the present study.

Compd. No.	X	R^1	R^2	R^3	R^4	Obs.	Model 3	Model 4
							pCC_{50} Est.	
1	O	CH_2CH_2OH	Me	H	Me	2·623	2·439	2·460
2	O	CH_2CH_2OH	Et	H	Me	2·258	2·328	2·304
3	O	CH_2CH_2OH	*t*-Bu	H	Me	1·875	2·108	2·016
4	O	CH_2CH_2OH	CH_2OH	H	Me	2·465	2·473	2·447
5	O	CH_2CH_2OH	CF_3	H	Me	2·292	2·210	2·118
6	O	CH_2CH_2OH	F	H	Me	2·450	2·321	2·286
7	O	CH_2CH_2OH	Cl	H	Me	2·322	2·294	2·272
8	O	CH_2CH_2OH	Br	H	Me	2·149	2·271	2·260
9	O	CH_2CH_2OH	I	H	Me	2·025	2·201	2·225
10	O	CH_2CH_2OH	OH	H	Me	2·230	2·516	2·483
11	O	CH_2CH_2OH	Me	Me	Me	2·649	2·500	2·539
12	O	CH_2CH_2OH	Me	Me	Me	2·386	2·329	2·331
13	O	CH_2CH_2OH	COOMe	H	Me	2·114	1·875	2·042
14	O	CH_2CH_2OH	COMe	H	Me	2·236	2·212	2·272
15	O	CH_2CH_2OH	COOH	H	Me	2·344	2·255	2·266
16	O	CH_2CH_2OH	CN	H	Me	2·358	2·372	2·345
17	O	CH_2CH_2OH	H	H	Allyl	2·547	2·375	2·396
18	S	CH_2CH_2OH	H	H	Et	2·369	2·328	2·408
19	S	CH_2CH_2OH	H	H	Pr	2·262	2·415	2·402
20	S	CH_2CH_2OH	ME	Me	*i*-Pr	2·170	2·355	2·391
21	S	CH_2CH_2OH	Cl	Cl	Et	2·362	2·245	2·246
22	O	CH_2CH_2OH	H	H	Et	1·716	2·025	2·010
23	O	CH_2CH_2OH	H	H	Pr	1·806	1·988	2·000
24	O	CH_2CH_2OH	H	H	*i*-Pr	2·602	2·438	2·433
25	O	CH_2CH_2OH	Me	Me	Et	2·387	2·328	2·288
26	O	CH_2CH_2OH	Me	Me	*i*-Pr	2·364	2·328	2·309
27	O	CH_2CH_2OH	Cl	Cl	Et	2·173	2·219	2·176
28	O	CH_2CH_2OH	H	H	H	2·107	2·108	2·053
29	O	$CH_2CH_2OnC_5H_{11}$	H	H	Me	2·871	2·658	2·732
30	O	$CH_2CH_2OCH_2Ph$	H	H	Me	1·740	1·561	1·608
31	O	Me	H	H	Me	1·653	1·491	1·739
32	O	Et	H	H	Me	2·387	2·282	2·357
33	O	Pr	H	H	Me	2·364	2·172	2·197
34	O	Bu	H	H	Me	2·167	2·092	2·068
35	O	CH_2Ph	H	H	Me	1·919	1·951	1·908
36	S	Et	H	H	Et	1·978	1·770	1·891
37	S	Et	Cl	Cl	Et	1·908	1·978	1·999
38	O	Et	H	H	Et	1·653	1·611	1·609
39	O	Et	Cl	Cl	Et	1·663	1·942	1·885
40	O	*i*-Pr	H	H	Et	2·207	2·061	2·042
41	O	CH_2Ph	H	H	Et	1·653	1·691	1·651
42	O	CH_2CH_2Ph	H	H	Et	2·155	1·951	1·916
43	O	Et	H	H	*i*-Pr	1·230	1·583	1·355
44	O	Et	H	H	*i*-Pr	1·531	1·659	1·739
45	O	$R^| = H$	H	H	Me	1·580	1·549	1·418
46	O	$R^| = Me$	H	H	Me	2·025	1·925	1·905
47	O	$R^| = Et$	H	H	Me	2·398	2·567	2·671
48	O	$R^| = u$	H	H	Me	1·949	2·322	2·198
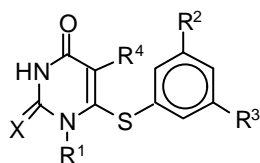
Figure 1. HEPT analogues used in the present study.

2.2 *Regression analysis*

We have used the maximum R^2 improvement as well as leave-one-out methods to identify prediction models.[17–19] When using the latter, there is no need to divide the compounds into training and test sets. The method of maximum R^2 finds the "best" one variable model, the "best" two variable model and so forth for the prediction of property/activity. Several models (combinations of variables) were examined to identify combinations of variables with good prediction capabilities. In all regression models developed, we have examined a variety of statistics associated with residues, i.e. the Wilks–Shapiro test for normality and Cooks D-statistics for outliers, to obtain the most reliable results.[18,19] Finally, we have used leave-one-out method and cross-validation for the investigation of the predictive power of the proposed models.[18–20]

Multiple regression analyses for correlating anti-HIV activities of the present set of compounds with the aforementioned molecular descriptors were carried out using *Regress-1* software as supplied by Professor I Lukovits, Hungarian Academy of Sciences, Budapest. Several multiple regressions were attempted using the correlation matrix from this program and the best results are considered and discussed in developing QSAR and, hence, for modelling the anti-HIV activities of the compounds in the present study.

2.3 *Indicator parameters (Ip_1, Ip_2, Ip_3)*

The indicator parameters[18,21] (variables) take on only two values, usually zero and one. The two values signify that the observation falls in one of two possible categories. The numerical values of the dummy variables are not intended to reflect a quantitative ordering of categories, but only serve to identify category or class membership. Therefore, they show the significance of a particular group or a substituent in a given series of drug. They account

for the abrupt increase or decrease of a given pharmacological activity at any specific site in the drug molecule. If the coefficient of indicator parameter carries a negative sign in the regression expression, this makes it very clear that the compound having this particular group at a particular position has considerably lower potency.

In the present case, the indicator parameter Ip_I is taken as 1 when the $-CH_2CH_2OH$ moiety is present at R^1 otherwise it is zero. When halogen is present at R^2 then indicator parameter Ip_2 is used whose value is taken as unity. For the presence of Me at R^4 the indicator parameter used is Ip_3 whose value is taken as unity.

2.4 *Equalized electronegativity (c_{eq})*

The equalized electronegativity (c_{eq})[11–15] has been calculated using the formula,

$$c_{eq} = N/\Sigma v/x, \tag{1}$$

where, $N = \Sigma v$ = total no. of atoms in the species, v = no. of atoms of a particular element in the species, and, x = electronegativity of that particular element. whereas, group electronegativity is defined as,

$$X_G = N_G/(v/x), \tag{2}$$

where N_G = no. of atoms in the group formula.

2.5 *van der Waals volume (Vw)*

The van der Waals volumes of the various compounds used in the present study have been calculated using the method suggested by Morriguchi *et al*[16].

3. **Results and discussion**

Table 1 records the structural details of RT-inhibitors under present study. This table 1 also records observed and estimated values of cytotoxicity in pCC_{50} units. The calculated molecular descriptors (c_{eq}, Vw) and assumed indicator parameters (Ip_1, Ip_2, Ip_3) are presented in table 2.

The most straightforward way to perform QSAR analysis is to divide the set of molecules into training set and test set. We obtain the QSAR equations using the training set and then apply them on the

test set. Such an analysis clearly gauges the reliability of QSAR equations. However, if leave-one-out methodology is used and then cross-validation is done, there is no need for such division into training and test sets.

Table 2. Molecular descriptors of compounds used in the present study (see table 1).

Compd. No	c_{eq}	Vw	Ip_1	Ip_2	Ip_3
1	2·3839	2·667	1	0	1
2	2·3716	2·822	1	0	1
3	2·3516	3·129	1	0	1
4	2·4025	2·620	1	0	1
5	2·4844	2·805	1	0	1
6	2·4342	2·650	1	0	1
7	2·4052	2·687	1	0	1
8	2·4171	2·720	1	0	1
9	2·3794	2·817	1	0	1
10	2·4896	2·582	1	0	1
11	2·3716	2·820	1	0	1
12	2·3566	2·983	1	0	1
13	2·4223	2·923	1	0	1
14	2·4048	2·761	1	0	1
15	2·4391	2·756	1	0	1
16	2·4228	2·822	1	0	1
17	2·3866	2·700	1	0	0
18	2·3677	2·784	1	0	0
19	2·3566	2·938	1	0	0
20	2·3387	3·244	1	0	0
21	2·4083	3·114	1	1	0
22	2·3838	2·668	1	0	0
23	2·3715	2·822	1	0	0
24	2·3715	2·822	1	0	0
25	2·3609	2·974	1	0	0
26	2·3516	3·128	1	0	0
27	2·4252	2·998	1	1	0
28	2·4158	2·361	1	0	0
29	2·3436	3·297	0	0	1
30	2·3714	3·395	0	0	1
31	2·3932	2·292	0	0	1
32	2·3690	2·446	0	0	1
33	2·3644	2·557	0	0	1
34	2·3534	2·754	0	0	1
35	2·3764	3·006	0	0	1
36	2·3445	2·716	0	0	0
37	2·3894	3·046	0	1	0
38	2·3644	2·600	0	0	0
39	2·4062	2·930	0	1	0
40	2·3534	2·754	0	0	0
41	2·3655	3·160	0	0	0
42	2·3558	3·314	0	0	0
43	2·3534	2·790	0	0	0
44	2·3677	2·650	0	0	0
45	2·4078	1·895	0	1	0
46	2·3866	2·057	0	1	0
47	2·3698	2·211	0	1	0
48	2·3445	2·237	0	1	0

The first step in proposing significant models is the selection of descriptors. This is achieved by obtaining correlation matrix as shown in table 3. The descriptors selected are those which exhibit correlation coefficient inferior to 0·5 and lead to high *F*-value during automatic linear regression modelling when processing the data corresponding to the 48 compounds used.

The inspection of the correlation matrix (table 3) indicates that none of the molecular descriptors used independently correlate significantly with pCC_{50}. The data, however, show that Vw is a better parameter for use in the regression analyses for obtaining statistically significant models.

Preliminary regression analysis has shown the non-existence of statistically significant mono-parametric models. This shows that statistically significant models are only possible through multiple regression analysis. The statistically significant multi-parametric models obtained are presented in table 4. The regression parameters and the quality of correlations of these models are given in table 5.

A perusal of tables 4 and 5 show that the performance of various 2, 3 and 4 parametric regression models shed much light on structure–activity relationships. Indicator parameters Ip_1 and Ip_2 play dominant roles in the exhibition of cytotoxic activity (pCC_{50}) of the HIV-inhibitors used. In view of this, we fix upon two such models (models 3 and 4, refer tables 4 and 5). These models are tri- and tetra-parametric respectively as given by the following,

$$pCC_{50} = -3\cdot9261 + 0\cdot7173\,(\pm 0\cdot0836)Vw \\ - (0\cdot4259\,(\pm 0\cdot0509)Ip_1 + 0\cdot1303(\pm 0\cdot0643)Ip_2, \tag{3}$$

$$n = 48, \quad Se = 0\cdot1700, \quad R = 0\cdot8750, \quad F = 47\cdot912, \quad Q = 5\cdot1470.$$

$$pCC_{50} = -0\cdot2116 + 0\cdot6741\,(\pm 0\cdot0832)Vw - 1\cdot4878 \\ (\pm 0\cdot7151))c_{eq} - 0\cdot3931 \quad (\pm 0\cdot0516)Ip_1 + 0\cdot1833 \\ (\pm 0\cdot0670)Ip_2, \tag{4}$$

$$n = 48, \quad Se = 0\cdot1639, \quad R = 0\cdot8872, \quad F = 39\cdot730, \quad Q = 5\cdot4130.$$

The van der Waals volume (Vw) is a steric parameter, its coefficient is positive in both the models. The positive value of Vw shows that pCC_{50} depends on the size of the fragments attached at R^1, R^2, R^4 and X. It can thus be stated that increase in the

Table 3. Correlation matrix showing inter-correlation of molecular descriptors and their correlation with activity.

|	$\mathbf{c}_{eq}$	Vw	pCC_{50}	Ip_1	Ip_2	Ip_3	Ip_4
$\mathbf{c}_{eq}$	1·0000						
Vw	−0·1659	1·0000					
pCC_{50}	−0·3782	0·6252	1·0000				
Ip_1	0·3315	0·0410	−0·5679	1·0000			
Ip_2	0·3795	0·1256	0·1250	0·1706	1·0000		
Ip_3	−0·4911	0·2496	0·3116	−0·1449	−0·1009	1·0000	
Ip_4	0·3819	−0·2006	−0·2657	−0·0191	−0·0716	0·3944	1·0000

Table 4. Statistically significant QSAR models for HIV-reverse transcriptase.

Model no.	Regression expressions
(1)	$pCC_{50} = -4\cdot1386 + 0\cdot7100\,(\pm 0\cdot1307)Vw$
(2)	$pCC_{50} = -3\cdot9695 + 0\cdot7377\,(\pm 0\cdot0858)Vw - 0\cdot4086\,(\pm 0\cdot0519)Ip_1$
(3)	$pCC_{50} = -3\cdot9261 + 0\cdot7173\,(\pm 0\cdot0836)Vw - (0\cdot4259\,(\pm 0\cdot0509)Ip_1 + 0\cdot1303\,(\pm 0\cdot0643)Ip_2$
(4)	$pCC_{50} = -0\cdot2116 + 0\cdot6741\,(\pm 0\cdot0832)Vw - 1\cdot4878\,(\pm 0\cdot7151)\,)\mathbf{c}_{eq} - 0\cdot3931\,(\pm 0\cdot0516)Ip_1 + 0\cdot1833\,(\pm 0\cdot0670)Ip_2$

Table 5. Quality of the models given in table 4.

Model no	No. of parameters	Se	R^2	R	F	Q
1	1	0·2680	0·3990	0·6252	29·524	2·3320
2	2	0·1758	0·7437	0·8624	65·295	4·9056
3	3	0·1700	0·7656	0·8750	47·912	5·1470
4	4	0·1639	0·7871	0·8872	39·730	5·4130

Se – standard error of estimation, R_A^2 – adjusted coefficient of determination, R – correlation coefficient, F – F-ratio, Q – quality factor[16] ($Q = R/Se$)
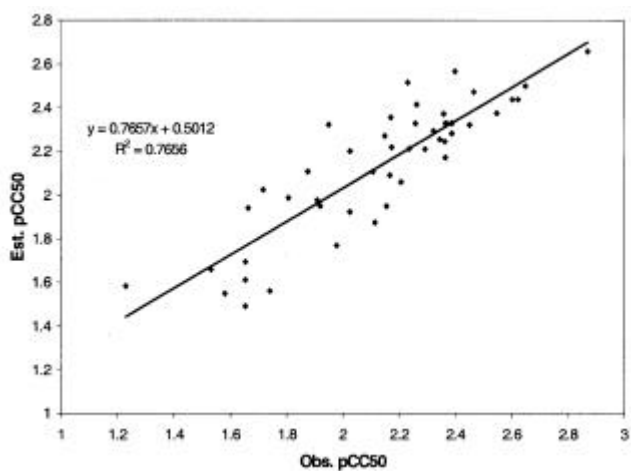
Figure 2. Correlation between observed and estimated pCC_{50} using model 3 (table 1).
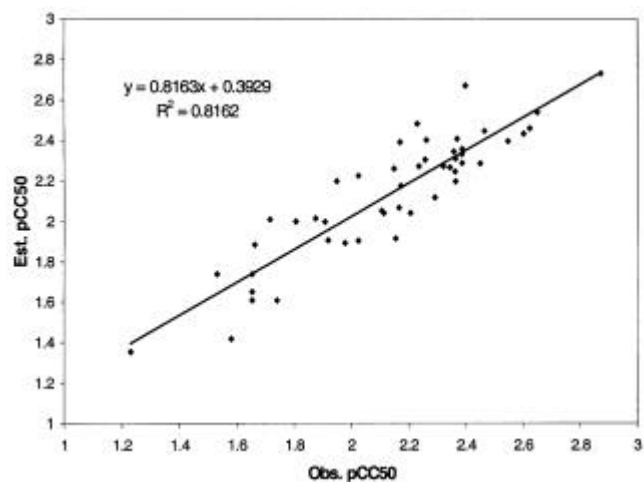
Figure 3. Correlation between observed and estimated pCC_{50} using model 4 (table 1).

Table 6. Cross-validation parameters[4–18] for the proposed models (table 4).

Model no	No. of parameters	Samples	PRESS	SSY	PRESS/ SSY	r_{cv}^2	S_{PRESS}	PSE
2	2	48	1·3902	4·0343	0·3446	0·6554	0·1758	0·1702
3	3	48	1·2713	4·1531	0·3061	0·6939	0·1699	0·1627
4	4	48	1·1551	4·2694	0·2706	0·7295	0·1639	0·1551

PRESS – predicted residual sum of squares, SSY – sum of the squares of regression value, r_{cv}^2 – cross-validation correlation coefficient, S_{PRESS} – uncertainty of prediction, PSE – predictive square error

bulk of the compounds likewise increase pCC_{50}. That is, for a compound to become a potent inhibitor R^1, R^2, R^4 and X positions should be occupied by bulky groups. Also, the magnitude of quality factor $(Q)^{22-24}$ indicates that the model expressed by (4) has better predictive potential.

Again, in the above models the indicator parameters, Ip_1 and Ip_2 are present. The former indicator parameter Ip_1 has a negative coefficient and is used when the –CH_2CH_2OH moiety is present at R^1 and the latter (Ip_2) has a coefficient which is positive and is used for the presence of halogen at R_2. Hence, presence of –CH_2CH_2OH is not helpful in exhibition of favourable cytotoxic activity (pCC_{50}).

The data presented in tables 4 and 5 also indicate that the quality of model 4 improves on introduction of an additional parameter, c_{eq}, the coefficient of which is negative.

This indicates that introduction of electron-withdrawing groups at R^1, R^2 and R^4 result in decrease in pCC_{50} value. Thus, highly electronegative groups at these positions are not preferred as it would inhibit the activity.

The predictive potential of the proposed models is determined by employing cross-validation method[14,15], and the cross-validation parameters estimated for the five proposed models are given in table 6. For the reasons given below these parameters suggest that both the models 3 and 4, (3) and (4) have better predictive potentials compared to others.

To be a reasonable QSAR model, PRESS/SSY should be the smaller than 0·4, and the value of this ratio smaller than 0·1 indicates an excellent model.[18] In the present case, except for models 1 and 4, all have this ratio smaller than 0·4.

It is interesting to record that the value of S_{PRESS} in the present case are very close to that of standard error of estimation, Se. Hence, both these parameters carry the same meaning and S_{PRESS} does not give any additional information regarding the uncertainty

of prediction. In view of this, we have calculated predictive square error, PSE,[18] as it seems to be most directly related to the uncertainty of prediction. The value of PSE (table 6) is found smallest for model 4 indicating that this model has the highest predictive potential.

Further confirmation of the predictive potential of the proposed models (3) and (4) is obtained by estimating predictive correlation coefficient. The $R_{pred}^2 = 0·7656$ (figure 2) and 0·8162 (figure 3) obtained for models 3 and 4, indicate that model 4 has the highest predictive potential.

4. Conclusions

On the basis of the above discussions we can draw the following conclusions:

(i) van der Waals volume (Vw) and equalized electronegativity (c_{eq}) are good parameters for modelling the cytotoxic activity (pCC_{50}) of the present set of compounds.

(ii) The introduction of electron-withdrawing groups at R^2 and R^4 results in decrease in cytotoxic activity (pCC_{50}). Thus, highly electronegative groups at these positions are not preferred.

References

1. Rizzo C R, Tirado-Rives J and Jorgensen W L 2001 *J. Med. Chem.* **44** 145

2. Yadav A and Singh S K 2003 *Bioorg. Med. Chem.* **11** 1801

3. Kavcher W and Devillers J (eds) 1990 *Practical applications of quantitative structure-activity relationships (QSAR) in environmental chemistry and toxicology* (Dordrect: Kluwer Academic)

4. Khadikar P V, Phadnis A and Shrivastava A 2002 *Bioorg. Med. Chem.* **10** 1181

5. Khadikar P V, Karmarkar S, Singh S and Shrivastava A 2002 *Bioorg. Med. Chem.* **10** 3163

6. Agrawal V K and Khadikar P V 2002 *Bioorg. Med. Chem.* **10** 3517

7. Karmarkar S, Agrawal V K, Mathur K C and Khadikar P V 2000 *Bulg. Chem. Ind.* **73** 99

8. Karmarkar S, Agrawal V K and Khadikar P V 2003 *Sci. Cult.* **69** 16

9. Diudea M V (ed.) 2000 *QSPR/QSAR studies by molecular descriptors* (Cluj, Romania: Babes-Bolyai University)

10. Khadikar P V, Mathur K C, Singh S, Phadnis A, Shrivastava A and Mandloi M 2002 *Bioorg. Med. Chem.* **10** 1761

11. Khadikar P V, Lukovits I, Agrawal V K, Shrivastava S, Jaiswal M, Gutman I, Karmarkar S and Srivastava A 2003 *Indian J. Chem.* **A42** 1436

12. Agrawal V K, Sohgaura R and Khadikar P V 2001 *Bioorg. Med. Chem.* **9** 3295

13. Agrawal V K, Mishra K and Khadikar P V 2003 *Oxid. Commun.* **26** 14

14. Pauling L 1969 *The nature of the chemical bond* (Ithaca, NY: Cornell Univ. Press)

15. Wells P R 1968 *Progress in physical organic chemistry* (New York: Interscience) vol. 6

16. Moriguchi I, Kanda Y and Komatsu K 1976 *Chem. Pharm. Bull.* **24** 1799

17. Tronchet J M J, Grigorov M, Dolatshahi N, Moriand F and Weber 1997 *J. Euro. J. Med. Chem.* **32** 279

18. Chaterjee S, Hadi S and Price B 2000 *Regression analysis by examples* 3rd edn (New York: Wiley)

19. Box G E B, Hunter W G and Hunter J S 1978 *Statistics for experiments* (New York: Wiley)

20. Lucic B and Trinajstic N 1999 *J. Chem. Inf. Comput. Sci.* **39** 121, 610

21. Trinajstic N 1992 *Chemical graph theory* 2nd edn (Boca Raton, FL: CRC Press)

22. Pogliani L 1994 *Amino Acids* **6** 141

23. Agrawal V K, Srivastava, R C and Khadikar P V 2001 *Acta Pharm.* **51** 117

24. Agrawal V K and Khadikar P V 2002 *Oxid. Commun.* **25** 184