Pergamon

Bioorganic & Medicinal Chemistry 10 (2002) 2919–2926

BIOORGANIC & MEDICINAL CHEMISTRY

QSAR Studies on Biological Activity of Piritrexim Analogues Against *pc* DHFR

Vijay K. Agrawal,[a] Rashmi Sohgaura[a] and Padmakar V. Khadikar[b,*]

[a]*Computer and QSAR Laboratories, A.P.S. University, Rewa-486003, India*
[b]*Research Division, Luxmi Fumigation and Pest Control, Pvt. Ltd., 3 Khatipura, Indore-452007, India*

Received 24 January 2002; accepted 23 April 2002

Abstract—Quantitative structure–activity relationship (QSAR) studies for 2,6-substituted 2,4-diaminopyrido[3,2-*d*] pyrimidine analogues of piritrexim (PTX) as inhibitors of dihydrofolate reductase (DHFR) are now made using topological indices. The results have shown that best models are obtained by multiparametric analysis. The predictive potential of the model is discussed on the basis of a cross-validation method.

Introduction

Recently, Gangjee et al.[1] have reported that piritrexim (PTX) (Fig. 1) is a lipophilic, new classical antifolate and is a potential inhibitor of dihydrofolate reductase (DHFR). It is currently in phase II clinical trials as an anticancer agent against a variety of tumors.[2–5] They have compared the structure of PTX with both, methotrexate (MTX) and trimetrexate (TMQ) and indicated that PTX differs from both these antitumor dihydrofolate reductase inhibitors, namely MTX and TMQ in that it contains a single side chain bridge rather than the usual two-atom bridge.

Consequent to this, Gangjee et al.[1] explored PTX analogues as potential inhibitors of dihydrofolate reductase (DHFR) and presented synthesis and biological activity of 21 6-substituted 2,4-diamino-pyridol [3,2-*d*] pyrimidine analogues of PTX (Fig. 2, Table 1) for this purpose. The potency of these PTX analogues as presented in Table 1 helped Gangjee et al.[1] to infer that the position of substitution influences inhibitory activity (IC_{50}) against pcDHFR. It is interesting to record that DHFR from *Pneumocytis carinii* (*pc*) is the target enzyme and is responsible for fatal opportunistic infections in AIDS patients.

The aforementioned substitution effect, however, failed to establish any quantitative structure–activity relationship (QSAR) in modelling inhibitory activity of referred PTX analogues. Consequent to this and as an extension of our earlier studies in attempting the role of the use of topological indices (TIs) for QSAR studies, we have

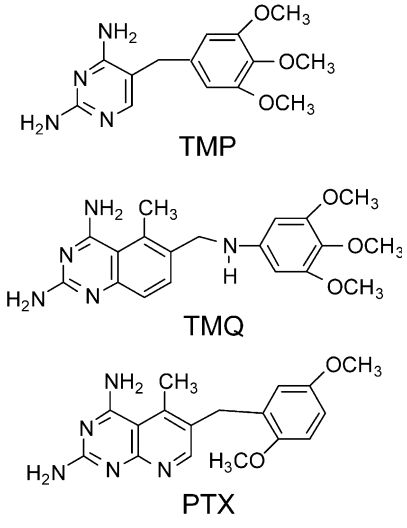

Figure 1. Structures of TMP, TMQ and PTX.

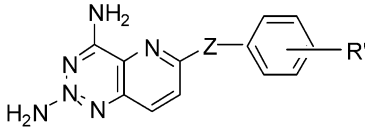

Figure 2. PTX analogues used in the present study.

*Corresponding author. Tel.: +91-731-531906; fax: +91-731-531906; e-mail: vijay-agrawal@lycos.com

0968-0896/02/$ - see front matter
PII: S0968-0896(02)00159-1

undertaken the present investigation in that we have used Wiener (W),[6] Szeged (Sz),[7,8] and molecular connectivity ($^0\chi^V$, $^1\chi^V$)[9,10] indices for this purpose. The results as discussed below show that these topological indices can be successfully used for modelling inhibitory activity of PTX analogues against *pc* DHFR. We have used simple as well as multilinear regressions for obtaining statistically significant models.[11,12] The predictive potential of the proposed models is finally investigated using a cross-validation method.[11,12]

The progress in chemical graph theory and topology has revealed their applicability to practically all areas of chemistry, particularly medicinal chemistry and pharmaceutical chemistry, mainly by successful predictions of physico-chemical properties and biological activities of organic molecules proceeding from their structure.[13–15] This indicated a possibility to explain why topological indices work so well. Such an attempt has been reported[16] for one of the most successful topological indices, the Wiener index.[6] Other widely used topological indices for this purpose are the valence connectivity indices ($^0\chi^V$, $^1\chi^V$).[9,10] Since the Wiener index is not applicable to a cycle containing compounds, Gutman has introduced the Szeged index[7,8] (Sz), which is considered as the modification of the Wiener index to cyclic compounds. The use of the Szeged index (Sz) in QSPR/QSAR is now well established.[17–20]

Randic[21] has shown that combinations of two or more topological indices result in more successful and statistically more significant models. Hence, another objective of the present study is to obtain such models and is an extension of our earlier work.[22–25] The results are presented and discussed in the following section.

Table 1. 6-Substituted 2,4-diaminopyrido[3.2.1] pyrimidine analogues of piritrexim used in the present study their pIC_{50} values against PcDHFR

Compd	*Z*	*R′*	pIC_{50}
1	S	2′-OMe	0.3424
2	S	4′-OMe	−0.1549
3	S	3′4′-diOMe	−1.0655
4	SO_2	2′-OMe	0.5051
5	SO_2	4′-OMe	1.02118
6	SO_2	3′4′-diOMe	0.4313
7	NH	2′-OMe	0.9395
8	NH	4′-OMe	1.9561
9	NH	3′4′-diOMe	1.6063
10	NCH_3	4′-Ome	−0.6575
11	NH	2′5′-diOMe	1.2068
12	NCH_3	2′5′-diOMe	−1.4685
13	NH	3′4′5′-triOMe	1.4132
14	NCH_3	3′4′5′-triOMe	−1.6777
15	NH	3′4′-C_4H_4	1.1760
16	NCH_3	3′4′-C_4H_4	−1.4989
17	NH	H	0.9190
18	NCH_3	H	−1.0021
19	NH	4′-Cl	1.1643

Results and Discussion

One-variable correlations

All the parameters described in the Experimental and presented in Tables 1 and 2 have been examined using a standard program for simple and multiple linear regression analyses with one, two, and three parameters (looking to size of the compounds used). The correlation matrix (Table 3) shows that none of the topological indices used results in any statistically significant one-variable model. However, Table 3 does show that the inhibitory activity (pIC_{50}) is significantly correlated with IP_1.

The correlation matrix (Table 3) further shows that high collinearity exists between: (i) W and Sz; (ii) W and $^0\chi^V$; and (iii) Sz and $^0\chi^V$. Comparatively less collinearity exists between: (i) W and $^1\chi^V$; (ii) Sz and $^1\chi^V$; and (iii) $^0\chi^V$ and $^1\chi^V$; and that both IP_1 and IP_2 do not correlate with any of the topological index used. However, compared to IP_1, the collinearity of IP_2 with topological indices, used is more pronounced.

Table 2. Calculated molecular descriptor and indicator parameters for 6-substituted, 24-diaminopyrido[3,2-*d*] pyrimidine analogous of piritrexim

Compd	W	Sz	$^0\chi^V$	$^1\chi^V$	IP_1	IP_2
1	1516	2454	11.2768	6.1774	0	1
2	1584	2590	11.2768	6.1715	0	1
3	1940	3134	12.6077	6.7004	0	0
4	1762	2820	12.0933	6.5107	0	1
5	1838	2972	12.0933	6.5048	0	1
6	2216	3546	13.4242	5.8839	0	0
7	1674	2708	11.3686	6.2692	1	1
8	1710	2780	11.3686	6.2633	1	1
9	2077	3339	12.6994	6.7918	1	0
10	1710	2780	12.3158	6.6576	0	1
11	2013	3211	12.6994	6.7922	1	0
12	2013	3211	13.6466	7.1866	0	0
13	2464	3926	14.0304	7.3213	1	0
14	2464	3926	14.9776	7.3824	0	0
15	1703	3197	12.1924	7.1449	1	0
16	1703	3197	13.1396	7.5393	0	0
17	1323	2169	10.0377	5.7402	1	0
18	1323	2169	10.9849	6.1346	0	0
19	1504	2462	10.3383	5.8399	1	0

The aforementioned results show that any two or more variable correlations involving any combination of W, Sz and $^0\chi^V$ will suffer from the defect of collinearity. Such a situation is discussed critically in the subsequent section in that it is shown that in spite of collinearity defect the proposed models are statistically significant.

Two-variable correlations

Because of the failure of obtaining statistically significant one-variable correlations, we attempted several two-variable correlations. All pairs of variables have been examined and only five such pairs were found to be statistically significant (Table 4). Among these five two-variable correlations; the correlation involving $^1\chi^V$ and IP_1 is found to be the best. This correlation is found as under:

$$\begin{aligned} p\ IC_{50}(\mu M) = &-0.6165(\pm 0.3015)\ {}^1\chi^V + 1.7096 \\ &\times (\pm 0.3251)IP_1 + 3.6080 \end{aligned} \quad (1)$$

This means that degree of unsaturation, the presence of hetero-atoms, contained in $^1\chi^V$ and presence of NH at Z position play dominant role in the exhibition of inhibitory activity of PTX analogues.

It is worth recording that only a pair of topological indices (W and $^0\chi^V$) gave a statistically significant model. However, its quality (Table 4) shows that it is less important than the two-variable model proposed above. However, statistics of the model [eq (1), Table 4] made it necessary for us to search for a better three-variable correlation.

Three-variable correlations

Out of the several three-variable correlations attempted only five were found to be statistically significant (Table 4). All these three-variable correlations were found to be statistically more significant than the best two-variable correlation discussed above [eq (1)]. However, among the five three-variable correlations, the correlation involving Sz, $^0\chi^V$, and IP_2 was found to be the best. This correlation is found as below:

$$\begin{aligned} pIC_{50}(\mu M) = &\ 0.0053(\pm\ 8.1955 \times 10^{-4})Sz \\ &- 2.2762(\pm 0.3233)\ {}^0\chi^V \\ &+ 0.5618(\pm 0.3105)IP_2 + 12.0758 \end{aligned} \quad (2)$$

Corresponding expressions contains $^1\chi^V$ and IP_1 in place of $^0\chi^V$ and IP_2, respectively were found statistically less significant. In this model [eq (2)] both Sz and IP_2play positive role. Record that IP_2 is used for the presence of OCH_3 at R′. Hence, the molecular shape, size and presence of $-OCH_3$ favors the activity.

Four-variable correlations

Looking to the size of the sample and the quality of the aforementioned three-variable correlation [eq (2)] makes the search of four-variable correlations necessary. In fact, seven such four-variable correlations were found statistically significant (Table 4). All these seven correlations were observed to be of better quality than the best three-variable correlation discussed above [eq (2)]. However, the correlation involving Sz, $^0\chi^V$, IP_1 and IP_2 gave best statistics. This correlation is found as under:

$$\begin{aligned} p\ IC_{50}(\mu M) = &\ 0.0034(\pm 0.0011)Sz \\ &- 1.4013(\pm 0.4499)\ {}^0\chi^V \\ &+ 0.9912(\pm 0.3998)IP_1 \\ &+ 0.7708(\pm 0.2809)IP_2 + 6.7073 \end{aligned} \quad (3)$$

The excellent statistics of eq (3) and looking again to sample size makes the search of five-variable correlations unnecessary. In fact, a five-variable correlation involving W, Sz, $^0\chi^V$, IP_1 and IP_2 and having better statistics has been obtained (Table 4). This correlation is found as under:

$$\begin{aligned} pIC_{50}(\mu M) = &\ 0.0018(\pm 0.0014)W \\ &+ 0.0024(\pm 0.0013)Sz \\ &- 1.4787(\pm 0.4440)\ {}^0\chi^V \\ &+ 0.9416(\pm 0.3928)IP_1 \\ &+ 0.6873(\pm 0.2822)IP_2 + 7.3351 \end{aligned} \quad (4)$$

Remark about collinearity defect

The three-, four-, and five-variable models [eqs (2), (3) and (4), respectively] suffer from the collinearity defect due to the involvement of topological indices having high collinearity (Table 3) and that such a defect is most pronounced in the five-variable model [eq (4)]. In spite of this collinearity defect, all the proposed correlations were found statistically significant. Such a situation was

Table 3. Correlation matrix for the inter-correlation of activity and molecular descriptors used in the present study

	pIC_{50}	W	Sz	$^0\chi^V$	$^1\chi^V$	IP_1	IP_2
pIC_{50}	1.00000						
W	−0.0582	1.00000					
Sz	−0.09167	0.96215	1.00000				
$^0\chi^V$	−0.38004	0.91234	0.93801	1.00000			
$^1\chi^V$	−0.36070	0.59507	0.72340	0.78541	1.00000		
IP_1	0.77325	−0.02453	−0.00771	−0.27566	−0.09489	1.00000	
IP_2	0.19783	−0.31601	−0.38469	−0.39414	−0.30799	−0.209361	1.00000

critically discussed by Randic[26] in that he proposed that selection of descriptors to be used in QSPR/QSAR studies should not be delegated solely to the computers although the statistical criteria will continue to be useful for preliminary screening of descriptors from a larger pool. Often in an automated selection of descriptors, a descriptor will be discarded because it is highly correlated with another descriptor already selected. What is important is not whether two descriptors parallel to one another, that is duplicate much of the same structural information but whether they in those parts that are important for QSPR/QSAR correlations. 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 property/activity then one of them can be discarded.

Considering the aforementioned recommendations of Randic[26] all the three correlations [eqs (2), (3) and (4)]

Table 4. Regression parameters and quality correlation of the significant regression equations

Compd	Parameter used	A_i I = 1,2,3,4,5	Standard deviation	Intercept B	Se	F	R^2_A	R	P	$Q = R/\mathrm{Se}$
1	W	-1.3772×10^{-4}	5.5780×10^{-4}	−0.2237	0.7806	11.972	0.5494	0.7742	6.617×10^{-4}	0.9918
	IP_1	1.7705	0.3628							
2	Sz	-1.9598×10^{-4}	3.59155×10^{-4}	0.1093	0.7749	12.267	0.5559	0.7780	5.895×10^{-4}	1.0040
	IP_1	1.7712	0.3601							
3	$^0\chi^V$	−0.1657	0.1455	1.6015	0.7522	13.509	0.5816	0.7925	3.663×10^{-4}	1.0535
	IP_1	1.6585	0.3636							
4	$^1\chi^V$	−0.6165	0.3015	3.6080	0.6964	17.096	0.6414	0.8254	1.066×10^{-4}	1.1852
	IP_1	1.7096	0.3251							
5	W	0.0061	0.0013	11.1528	0.7384	14.323	0.5968	0.8010	2.72×10^{-4}	1.0840
	$^0\chi^V$	−1.7903	0.3354							
6	W	0.0039	0.0013	6.9843	0.6154	16.422	0.7199	0.8756	5.266×10^{-5}	1.4220
	$^0\chi^V$	−1.1663	0.3558							
	IP_1	1.0320	0.3641							
7	Sz	0.0037	0.0013	9.3312	0.6221	15.964	0.7138	0.8726	6.175×10^{-5}	1.4026
	$^0\chi^V$	−1.6539	0.5276							
	IP_1	0.6619	0.4569							
8	Sz	0.0053	8.1955×10^{-4}	12.0758	0.6018	17.406	0.7322	0.8814	3.782×10^{-5}	1.4646
	$^0\chi^V$	−2.2762	0.3233							
	IP_2	0.5618	0.3105							
9	W	7.2578×10^{-4}	6.1173×10^{-4}	4.0184	0.6877	12.157	0.6503	0.8418	2.693×10^{-4}	1.2240
	$^1\chi^V$	−0.8784	0.3707							
	IP_1	1.6944	0.3212							
10	W	0.0026	0.0016	13.1496	0.6133	16.570	0.7218	0.8765	5.006×10^{-5}	1.4291
	Sz	0.0035	0.0012							
	$^0\chi^V$	−2.3012	0.3309							
11	W	0.0038	0.0022	12.2771	0.6201	12.326	0.7157	0.8825	1.670×10^{-4}	1.4231
	Sz	0.0029	0.0015							
	$^0\chi^V$	−2.5064	0.4176							
	$^1\chi^V$	0.4701	0.3726							
12	W	0.0033	0.0012	4.3972	0.5523	16.447	0.7744	0.9080	3.468×10^{-5}	1.6440
	$^0\chi^V$	−0.8901	0.3442							
	IP_1	1.3468	0.3581							
	IP_2	0.6592	0.3066							
13	W	9.6679×10^{-4}	5.1739×10^{-4}	2.1355	0.5733	15.017	0.7570	0.9005	5.754×10^{-5}	1.5707
	$^1\chi^V$	−0.7173	0.3145							
	IP_1	1.8840	0.2765							
	IP_2	0.8275	0.3004							
14	Sz	0.0034	0.0011	6.7073	0.5193	19.069	0.8006	0.9192	1.492×10^{-5}	1.7700
	$^0\chi^V$	−1.4013	0.4499							
	IP_1	0.9912	0.3998							
	IP_2	0.7708	0.2809							
15	Sz	8.6773×10^{-4}	3.7859×10^{-4}	2.7529	8.5464	0.9101	16.882	0.7792	2.993×10^{-5}	1.6637
	$^1\chi^V$	−0.9390	0.3451							
	IP_1	1.8648	0.2636							
	IP_2	0.8872	0.2902							
16	W	0.0025	0.0016	9.8254	0.5904	0.8942	13.959	0.7423	8.582×10^{-5}	1.5145
	Sz	0.0022	0.0015							
	$^0\chi^V$	−1.7259	0.5026							
	IP_1	0.6418	0.4338							
17	W	0.0021	0.0016	12.5010	0.5874	0.8953	14.134	0.7448	8.022×10^{-5}	1.5241
	Sz	0.0041	0.0012							
	$^0\chi^V$	−2.3161	0.3171							
	IP_2	0.4754	0.3101							
18	W	0.0018	0.0014	7.3351	0.5077	0.9286	16.290	0.8094	3.380×10^{-5}	1.8290
	Sz	0.0024	0.0013							
	$^0\chi^V$	−1.4787	0.4440							
	IP_1	0.9416	0.3928							
	IP_2	0.6873	0.2822							

in spite of the correlation defect be considered statistically significant as each of the parameters involved, in spite of their collinearity, have different information content. In our earlier communications[24,25] we have stated that though very high collinearity exists between W and Sz, Sz has some hither to unknown information content not present in W. Hence, we can have a statistically significant model, such as the model expressed by eq (4), in that both W and Sz are present simultaneously in the proposed model. Furthermore, IP_1 and IP_2 are used as dummy parameters to account for the structural feature not contained in the topological indices used.

Remark on the number of parameters

At this stage it is worth making remarks on the number of parameters involved in the proposed models, particularly on the parameters involved in the model expressed by eq (4). Fortunately a thumbs rule is available in the literature[25] with regards to maximum number of variables (molecular descriptors) to be used compared to the size of the sample (number of compounds used). This useful rule of thumb is that the number of compounds should be 3–6 times the number of parameters (molecular descriptors) used in the proposed model.

Considering the aforementioned thumb rule we observed that the model proposed by eqs (1)–(3) are much above the maximum range, while the model proposed by eq (4) is just nearer to the maximum range (five-molecular descriptor for the set of 20 compounds used) of the applicability of thumb rule.

The usefulness of eq (4) compared to eq (3) can be judged by estimating pIC_{50} values using each of these correlations and comparing these estimated pIC_{50} values with the corresponding observed values of pIC_{50}. Such a comparison is made in Table 5 and demonstrated in Figure 3. The predictive correlation potential ($r = 0.9285$ for model 4) indicate that model 4 ($r^2 = 0.862$) is a better model than model 3 ($r^2 = 0.845$).

More about the parameters involved

We now throw more light on the best two-, three-, four- and five-variable correlations proposed through eqs (2)–(4), respectively.

The first-order valence connectivity index, $^1\chi^V$, distinguish the degree of unsaturation and the presence of hetero-atoms; while the indicator parameter IP_1 is used for substitution at *Z*.

Hence, the model proposed by eq (1) is governed by the degree of unsaturation, presence of hetero-atoms and presence of NH substituent at Z and that it (IP_1) has a positive role in modelling inhibitory activity through the model proposed by eq (1).

As stated earlier, Sz accounts for the cyclicity of the compounds used. In addition to the number of cycles present in the molecule we have shown that tree-like (acyclic) side chain can also be accounted for the cycle containing compounds. In the set of PTX-analogues used the number of cycles involved in them is the same (i.e., three cycles). Hence, the variation in Sz will account for the combined effect due to variations caused by Z and R′ (Fig. 2). Also, $^0\chi^V$ accounts for zero-order connectivity and presence of hetero-atoms and that we have used IP_2 for the substitution of OCH_3 at R′. Hence, the number cycles, combined substitution pattern at Z, and R′; for zero-order connectivity, presence

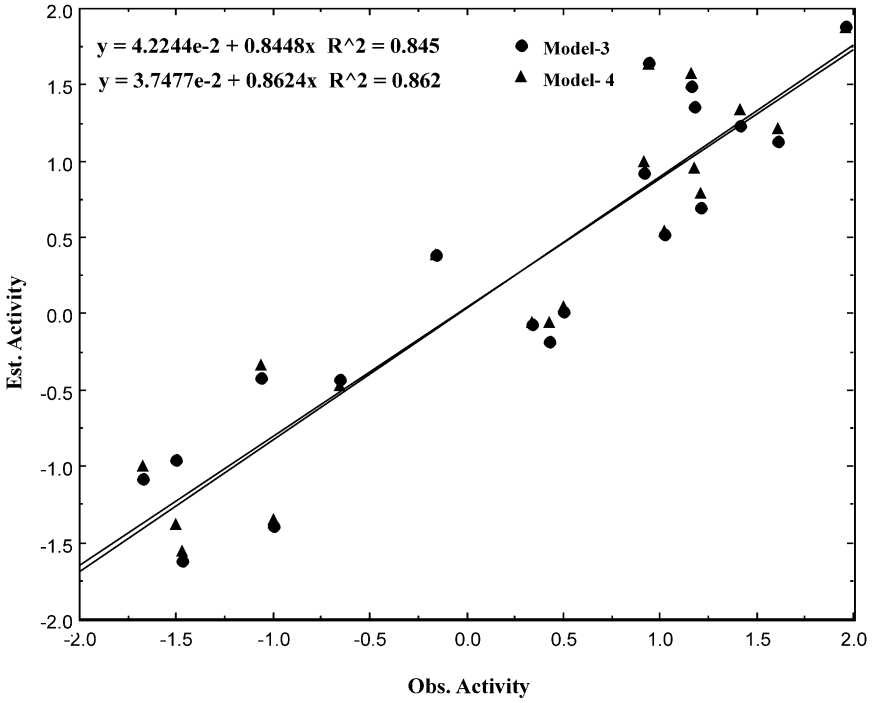

Figure 3. Correlation of observed versus estimated pIC_{50} using models 3 and 4, respectively.

of hetero-atoms and IP_2 is responsible for modelling inhibitory activity through model proposed by eq (2).

Inhibitory potential of PTX analogues through the model proposed vide eq (3) is governed by each of the characteristics of Sz, $^0\chi^V$, IP_1 and IP_2 mentioned above. In this model, the characteristics (mentioned above) associated with Sz, IP_1 and IP_2 play a positive role in the exhibition of inhibitory potential, while characteristics associated with $^0\chi^V$ play a negative role in the exhibition of log IC_{50}.

Finally, the model proposed through eq (4) accounts for the characteristics associated with W, Sz, $^0\chi^V$, IP_1 and IP_2. Here, the characteristics associated with Sz, $^0\chi^V$, IP_1 and IP_2 are already mentioned above. Here, W is also present in the model. The Wiener index (W) accounts for the size, shape and branching present in the molecules. Hence, the presence of W accounts for the effect of size, shape and branching on the inhibitory activity of PTX analogues used.

It is interesting to record that except the two-variable model [eq (1)] all other proposed models contain invariably $^0\chi^V$ as one of the correlating parameters. It means that compared to $^1\chi^V$, $^0\chi^V$ is more important in the exhibition of inhibitory potential of PTX analogues. This is also supported by the correlation matrix (Table 3) in that correlation between pIC_{50} and $^0\chi^V$ (−0.3800) is comparatively more than the correlation between pIC_{50} and $^1\chi^V$ (−0.3607).

It is worth mentioning that the correlation potential of the proposed model should be confirmed on the basis of combined effects of correlation coefficient (R) and standard error of estimation (Se). This combined effect can be discussed on the basis of quality factor Q,[27,28] which is defined as the ratio of correlation coefficient (R) to the standard error of estimation (Se), namely $Q = R/Se$. That is, the higher the magnitude of R, the lower the Se, the bigger will be Q, and the best will be the quality of proposed model. These Q values for the different correlations attempted are presented in Table 4 showing that the model proposed vide eq (4) has the best correlation potential.

Remark on predicting potential

It is worth mentioning that the models having the best correlation potential need not have the best predictive value too. As opposed to traditional regression methods, the cross-validation evaluates the validity of a model by how well it predicts data rather than how well it fits data. The analysis uses a 'leave-one-out' scheme, a model is built with $N-1$ compounds and the Nth compound is predicted. Each compound is left out of the model derivation and predicted in turn. As indication of the performance of the model is obtained from the cross-validated (or predictive) r^2_{CV} which is defined as:

$$r^2_{CV} = \frac{SD - PRESS}{SD}$$

Table 5. Comparison of experimental and calculated inhibition pIC_{50} (μM) using models 3 and 4

Compd	Obs.	Estimated pIC_{50} using			
	pIC_{50}	Model-3	Res.	Model-4	Res.
1	0.342	−0.072	0.414	−0.062	0.404
2	−0.155	0.385	−0.540	0.385	−0.540
3	−1.066	−0.421	−0.645	−0.331	−0.735
4	0.505	0.014	0.491	0.047	0.458
5	1.021	0.525	0.496	0.547	0.474
6	0.431	−0.180	0.611	−0.058	0.489
7	0.940	1.644	−0.704	1.634	−0.694
8	1.956	1.887	0.069	1.872	0.084
9	1.606	1.131	0.475	1.211	0.395
10	−0.658	−0.432	−0.226	−0.471	−0.187
11	1.207	0.700	0.507	0.790	0.417
12	−1.469	−1.618	0.149	−1.552	0.083
13	1.413	1.239	0.174	1.341	0.072
14	−1.678	−1.079	−0.599	−1.001	−0.677
15	1.176	1.364	−0.188	0.959	0.217
16	−1.499	−0.955	−0.544	−1.384	−0.115
17	0.919	0.926	−0.007	0.997	−0.078
18	−1.002	−1.392	0.390	−1.346	0.0344
19	1.164	1.490	−0.326	1.578	−0.414

Table 6. Cross-validation parameters for the proposed model

Model no.	Parameter involved	PRESS	SD	PRESS/SD	r^2_{cv}	S_{PRESS}	PSE	R^2_A
1	$^1\chi^V$, IP_1	7.7595	16.5818	0.4579	0.5320	0.6364	0.6390	0.6414
2	Sz, $^0\chi^V$, IP_2	5.4319	18.9094	0.2373	0.7127	0.6018	0.5347	0.7322
3	Sz, $^0\chi^V$, IP_1, IP_2	3.7749	20.5605	0.1836	0.8164	0.5193	0.4457	0.8006
4	W, Sz, $^0\chi^V$, IP_1, IP_2	3.3504	20.9909	0.1596	0.8404	0.5077	0.4199	0.8094

where SD is the sum-of-squares deviation for each activity from the mean. The predictive sum-of-squares (PRESS) is the sum of the squared difference between the actual and that predicted when the compound is omitted from the fitting process. Once a model is developed which has highest cross-validated r^2, this method is used to derive the conventional QSAR equation and conventional r^2 and s values. The results of the final model are often visualized as contour maps of the coefficient.

In addition to PRESS, SD, r^2_{cv}, S_{PRESS}, one also needs to evaluate predictive-square-error (PSE) in an attempt to decide the predictive potential of the proposed models.

These cross-validated parameters, namely PRESS, SD, r^2_{cv}, S_{PRESS} and PSE were calculated for each of the proposed models mentioned above and are presented in Table 6. In addition, Table 6 also records adjustable R^2 values (R^2_A) for the referred model.

A perusal of Table 6 shows that PRESS $<$SD for all the proposed model and that PRESS/SD ranges between 0.4579 and 0.1596. Hence, all the proposed models predict better than chance and can be considered 'statistically significant'. Except for the model proposed by eq (1), all other models are statistically significant and the model expressed by eq (4) is an excellent model. r^2_{cv} Values also support this finding. The order of r^2_{cv} values indicates overall predictive ability of the models: the smaller the r^2_{cv}, the better is the predictive ability.

Table 6 shows that uncertainty of prediction (S_{PRESS}) is the same as standard error of estimation (Se). Therefore, S_{PRESS} is not a better parameter for estimating uncertainty of prediction.

In view of the above, we have used PSE for accounting uncertainty of prediction. It seems that, compared to S_{PRESS}, PSE is more directly related to the uncertainty of predictions. The smaller the PSE value the better will be the prediction potential of the model. Hence, based on PSE value, we again observed that the model expressed by eq (4) has the best predictive ability.

Finally, a perusal of Table 6 shows that R^2_A value goes on increasing as we shift from bi- to penta-parametric models. It is particularly important when the number of independent variables is large relative to the sample size. In our case it is important for the model proposed by eq (4). R^2_A will decrease if the added variable does not have a significant contribution. In our case, as we see, that R^2_A goes on increasing as we pass from bi- to penta-parametric model indicating that all these models are statistically significant.

Conclusions

Reviewing the analysis of the best, one-, two-, three-, four- and five-variable regressions we arrive at the conclusion that even a five-variable regression model has excellent correlation potential. The cross-validation method further proposes that such models do have the best predictive ability. The results also show that inclusion of collinearity related parameters results in statistically significant multiparameteric models. Once again, here also, we obtained a multiparameteric model in that W and Sz are involved simultaneously and are still statistically significant. Such a result again proposes that in addition to shape, size and branching effect, Szeged index has some hitherto unknown information content. Research in this direction is under way and will be published soon.

Experimental

Inhibition potential

Inhibitory activity as reported by Gangjee et al.[1] as IC_{50} were converted into their log units (log IC_{50}) and used in the present investigation and finally used as pIC_{50}.

Calculation of topological indices

The calculations of topological indices (W, Sz, $^0\chi^V$, $^1\chi^V$) were made using the program provided by Prof. Istvan Lukovits, Hungarian Academy of Sciences, Budapest, Hungry. The following expressions are used in developing this program.

The molecular connectivity indices

The connectivity index $\chi = \chi(G)$ of a graph G is defined by Randic[21] as under

$$\chi = \chi(G) = \sum_{ij} [d_i \; d_j]^{-0.5} \tag{5}$$

where d_i is the valence of a vertex i, equal to the number of bonds connected to the atom i, in G, representing the graph of a compound. Meaning of d_j is analogous.

In the case of hetero-systems, the connectivity is given in terms of valence delta values δ^v_i and δ^v_j of atoms i and j and is denoted by χ^v. This version of the connectivity index is called the valence connectivity index and defined as[9,10] under:

$$\chi^v = \chi^v(G) = \sum_{ij} \left[\delta^v_i \; \delta^v_j\right]^{-0.5} \tag{6}$$

where the sum is taken over all bonds i–j of the molecule. Valence delta values are given by

$$\delta^v_i = \frac{Z^v_i - H_i}{Z_i - Z_j - 1} \tag{7}$$

where Z_i is the atomic number of atom i, Z^v_i is the number of valence electron of the atom i and H_i is the number of hydrogen atoms attached to atom i. the δ^v_i - values are available in the books of Kier and Hall.[9,10]

Recall that now-a-days the connectivity and the valence connectivity indices expressed by eqs (5) and (6) are

termed as first-order connectivity and first-order valence connectivity index, respectively. Lower or higher order indices are also possible which are defined analogously.

Wiener index (W)

The Wiener index (W) is the oldest topological index.[6] It is based on the vertex distances of the respective molecular graph and is defined as below.

Let us denote a molecular graph G and v_1, v_2, v_3,...,v_n its vertices. Let $d(v_l, v_| \mid G)$ stand for the distance between the vertices v_i and v_j. Then the Wiener index is defined as:

$$W = W(G) = 1/2\sum_{i=1}^{n}\sum_{j=1}^{n} d(v_i, v_j \mid G) \quad (8)$$

Szeged index (Sz)

Let e be an edge of the molecular graph G. Let $n_1(e \mid G)$ be the number of vertices of G lying closer to one end of e; let $n_2(e \mid G)$ be the number of vertices of G lying closer to the other end of e. Then the Szeged index (Sz) is defined[7,8] as:

$$\mathrm{Sz}(G) = \mathrm{Sz} = \sum_{e} n_1(e|G \mid n_2(e \mid G) \quad (9)$$

with the summation giving over all edges of G.

In cyclic graphs, there are edges equidistant from both the ends of edge e; by definition of Sz, such edges are not taken into account.

Indicator parameters (Ip_1, Ip_2)

Indicator variables (parameters), sometimes called dummy variables or de novo constants,[11] are used in linear multiple regression analysis to account for certain features which cannot be described by continuous variables. In QSAR equations, they normally describe a certain structural element, be it a substituent or another molecular fragment. Thus, Free Wilson[11] analysis may be interpreted as a regression analysis approach using only indicator variables.

Regression analysis

QSAR models were proposed using multiple regression analysis[14] and predictive potential of the models was determined by cross-validation methods.[15,16] Multiple regression analyses were carried out using Regress-1 software as supplied by Professor I. Lukovits, Hungarian Academy of Sciences, Budapest, Hungary. Similarly, W and Sz were calculated employing his Wiener-1 program.

Computation. All the computations were carried out in Power Macintosh 9600/233.

Acknowledgements

One of the authors (P.V.K.) is thankful to Professor Ivan Gutman for introducing him to the fascinating field of Chemical Graph Theory and Topology. The authors are thankful to Professor I. Lukovits, Hungarian Academy of Sciences, Budapest, Hungary for providing software.

References and Notes

1. Gangjee, A.; Zhu, Y.; Queener, S. F. *J. Med. Chem.* **1998**, *41*, 4533.
2. Grivsky, E. M.; Lee, S.; Sirel, C. W.; Duch, D. S.; Nichol, C. A. *J. Med. Chem.* **1980**, *23*, 327.
3. de Vries, E. G. E.; Gietema, J. A.; Workman, P.; Scott, J. F.; Crawshaw, A.; Dobbs, H. J.; Dennis, I.; Mulder, N. H.; Sleijiter, D. R.; Willemsen, P. H. B. *Br. J. Cancer* **1993**, *68*, 641.
4. de Wit, R.; Kaye, S. B.; Roberts, J. T.; Stoter, G.; Scott, J.; Verweij, J. *Br. J. Cancer*. **1993**, *67*, 388.
5. Gangjee, A.; Elzein, E.; Kothari, M.; Vasudevan, A. *Curr. Pharm. Res*. **1996**, *2*, 263.
6. Wiener, H. *J. Am. Chem. Soc*. **1947**, *69*, 17.
7. Gutman, I. *Graph Theory Notes New York* **1994**, *27*, 9.
8. Khadikar, P. V.; Deshpande, N. V.; Kale, P. P.; Dobrynin, A.; Gutman, I.; Domotor, G. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 547.
9. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic: New York, 1976.
10. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Relationship*; Wiley: New York, 1976.
11. Chaterjee, S.; Hadi, A. S.; Price, B. *Regression Analysis by Examples*, 3rd ed.; Wiley: New York, 2000.
12. Lucic, B.; Trinajstic, N. *J. Chem. Inf. Comput. Sci*. **1999**, 39, 121.
13. Roberto Toteschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
14. Diudea, M. V., Ed. *QSPR/QSAR Studies by Molecular Descriptors*; Babes-Bolyai University: Cluj, Romania, 2000.
15. King, R. B. *Chemical Applications of Topology and Graph Theory*; Elsevier: Amsterdam, 1983.
16. Bonchev, D. *Introduction Theoretical Indices for Characterization of Chemical Structure*; Research Studies: Chichesleu, UK, 1983.
17. Khadikar, P. V.; Kale, P. P.; Deshpande, N. V.; Agrawal, V. K. *J. Ind. Chem. Soc*. **2000**, *77*, 449.
18. Khadikar, P. V.; Karmarkar, S.; Agrawal, V. K. *Nat. Acad. Sci. Lett*. **2000**, *23*, 165.
19. Khadikar, P. V.; Kale, P. P.; Deshpande, N. V.; Karmarkar, S.; Agrawal, V. K. *MATCH* **2001**, *43*, 7.
20. Mandloi, M.; Agrawal, V. K.; Mathur, K. C.; Karmarkar, S.; Khadikar, P. V. *Acta. Pharma*. **2000**, *50*, 303.
21. Randic, M. *J. Am. Chem. Soc*. **1975**, *97*, 6609.
22. Agrawal, V. K.; Khadikar, P. V. *Bioorg. Med. Chem.* **2001**, *9*, 2787.
23. Agrawal, V. K.; Srivastava, R.; Khadikar, P. V. *Bioorg. Med. Chem*. **2001**, *9*, 3287.
24. Agrawal, V. K.; Sohgaura, R.; Khadikar, P. V. *Bioorg. Med. Chem*. **2001**, *9*, 3295.
25. Agrawal, V. K.; Khadikar, P. V. *Bioorg. Med. Chem.* **2001**, *9*, 3037.
26. Randic, M. *J. Am. Chem. Soc*. **1975**, *97*, 6609.
27. Pogliani, L. *Amino Acids* **1994**, *6*, 141.
28. Pogliani, L. *J. Phys. Chem*. **1996**, *100*, 18065.