CROATICA CHEMICA ACTA CCACAA **74** (2) 237–250 (2001)

ISSN-0011-1643
CCA-2728 *Original Scientific Paper*

Predicting Inhibition of Microsomal *p*-Hydroxylation of Aniline by Aliphatic Alcohols: A QSAR Approach Based on the Weighted Path Numbers

Dragan Amić,[a,*] *Bono Lučić,*[b,*] *Sonja Nikolić,*[b] *and Nenad Trinajstić*[b]

[a]*Faculty of Agriculture, The Josip Juraj Strossmayer University, 31001 Osijek, Croatia*

[b]*The Rugjer Bošković Institute, P. O. Box 180, HR–10002 Zagreb, Croatia*

Received January 2, 2001; revised February 16, 2001; accepted February 19, 2001

Weighted path numbers are used to build QSAR models for predicting inhibition of microsomal *p*-hydroxylation of aniline by aliphatic alcohols. Models with two, three and four weighted path numbers are considered. Fit and cross-validated statistical parameters are used to measure the model quality. The best statistical parameters possess models with four weighted path numbers. Comparison with models from the literature favors models based on the weighted path numbers.

Key words: aliphatic alcohols, inhibition of microsomal *p*-hydroxylation of aniline, QSAR, weighted path numbers.

INTRODUCTION

In this paper we report the QSAR[1] modeling of the inhibitory effects ($1/I_{50}$) of 19 aliphatic alcohols on microsomal *p*-hydroxylation of anilines using the recently proposed molecular descriptors named the weighted path numbers.[2–4] We have been interested in the physicochemical and biological properties of alcohols for some time because they are toxic materials and thus are, among other things, also dangerous environmental pollutants. Our previous studies were centered on the various aspects of the aqueous

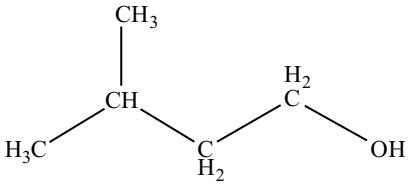
* Author to whom correspondence should be addressed.
(E-mail: D. A. – damic@suncokret.pfos.hr; B. L. – lucic@faust.irb.hr)

solubility of aliphatic alcohols[5–7] because the toxic action of alcohols on higher organisms, presumably through the phenomenom of narcosis,[8,9] depends on their solubility in water. Alcohols are also technologically important materials and are used in the manufacture of a number of products. Several of the QSAR models that we used to predict the aqueous solubility of aliphatic alcohols were based on the weighted path numbers.[10] These models possessed better statistical characteristics than any other model reported in the literature. Randić and Basak[2,4] also found that the structure-boiling point models of aliphatic alcohols based on the weighted path numbers are superior to models based on other kinds of molecular descriptors. Since the use of the weighted path numbers in the structure-property modeling of alcohols was fairly successful, we decided to use these descriptors in the structure-activity modeling of alcohols, employing the biological data of Cohen and Mannering.[11]

WEIGHTED PATH NUMBERS

Aliphatic alcohols will be represented by weighted trees.[12] As an example, a weighted tree, ^{w}T, representing the hydrogen-depleted skeleton of 3-methyl-1-butanol is shown in Figure 1.

(i) 3-methyl-1-butanol



(ii) Labeled hydrogen-depleted weighted tree

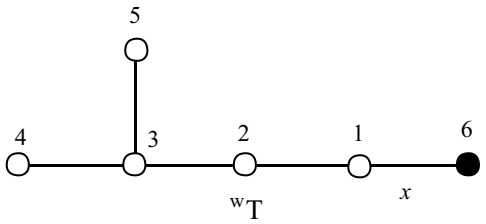


Figure 1. A labeled weighted tree, ^{w}T, representing the hydrogen-depleted skeleton of 3-methyl-1-butanol. The black dot marks the position of oxygen and x stands for the weight of C–O bond. Weights of C–C bonds are taken to be 1.

A path P_l (l is the length of the path) is a sequence of adjacent edges which do not pass through the same vertex more than once.[13] Note that P_1 is equal to the number of edges in the graph. A weighted path contains weighted edges. In the case of aliphatic alcohols, the weight of the edge representing C–O bond is taken to be x, while the weights of edges representing C–C bonds are all equal to 1 (see Figure 1). The count of paths and weighted paths in the weighted tree ^{w}T, given in Figure 1, is illustrated in Table I. The length of a path containing the edge with weight x is simply denoted by x.

TABLE I

The count of paths and weighted paths in a labeled weighted tree ^{w}T (see Figure 1), representing the hydrogen-depleted skeleton of 3-methyl-1-butanol

Vertex	P_1	P_2	P_3	P_4
1	$1 + x$	1	2	
2	2	$2 + x$		
3	3	1	x	
4	1	2	1	x
5	1	2	1	x
6	x	x	x	$2x$
Weighted path numbers	$4 + x$	$4 + x$	$2 + x$	$2x$

The weighted path numbers for 3-methyl-1-butanol are equal to the sums of weighted paths over all vertices in a graph divided by 2, since each vertex is counted twice. If we assume $x = 1$, then the count ($P_1 = 5$, $P_2 = 5$ $P_3 = 3$, $P_4 = 2$) represents paths in 2-methylpentane. If x is left undefined, then it can be viewed as a variable to be adjusted from regression analysis to obtain the smallest standard error of estimate for the regression considered. This is the idea underlying the use of variable (optimal) molecular descriptors in the structure-property-activity modeling.[2–4,14,15]

RESULTS AND DISCUSSION

In Table II, we give the weighted path numbers and log values of experimental values $\log(1/I_{50})$ for inhibition of microsomal cytochrome P450 *p*-hydroxylation of anilines by 19 aliphatic alcohols, taken from Cohen and Mannering.[11] The inhibitory effect of aliphatic alcohols on aniline

TABLE II

The weighted path numbers P_l (l = 1,2,...,7) and the observed inhibition of microsomal *p*-hydroxylation of anilines log ($1/I_{50}$) by aliphatic alcohols

Aliphatic alcohol	P_1	P_2	P_3	P_4	P_5	P_6	P_7	log ($1/I_{50}$)
Ethanol	$1+x$	x						–1.0969
1-Propanol	$2+x$	$1+x$	x					–0.4771
1-Butanol	$3+x$	$2+x$	$1+x$	x				–0.0453
1-Pentanol	$4+x$	$3+x$	$2+x$	$1+x$	x			0.2676
1-Hexanol	$5+x$	$4+x$	$3+x$	$2+x$	$1+x$	x		0.5376
1-Heptanol	$6+x$	$5+x$	$4+x$	$3+x$	$2+x$	$1+x$	x	0.6778
2-Methyl-1-propanol	$3+x$	$3+x$	$2+x$					–0.3874
2-Methyl-1-butanol	$4+x$	$4+x$	$2+2x$	x				–0.1461
3-Methyl-1-butanol	$4+x$	$4+x$	$2+x$	$2x$				–0.1931
2,2-Dimethyl-1-propanol	$4+x$	$6+x$	$3x$					–0.6702
2-Propanol	$2+x$	$1+2x$						–0.4713
2-Butanol	$3+x$	$2+2x$	$1+x$					–0.3483
2-Pentanol	$4+x$	$3+2x$	$2+x$	$1+x$				–0.0719
2-Hexanol	$5+x$	$4+2x$	$3+x$	$2+x$	$1+x$			0.1484
2-Heptanol	$6+x$	$5+2x$	$4+x$	$3+x$	$2+x$	$1+x$		0.2518
3-Pentanol	$4+x$	$3+2x$	$2+2x$	1				–0.3655
3-Hexanol	$5+x$	$4+2x$	$3+2x$	$2+x$				–0.4698
2-Methyl-3-pentanol	$5+x$	$5+2x$	$3+3x$	2				–0.8910
2,4-Dimethyl-3-pentanol	$6+x$	$7+2x$	$4+4x$	4				–1.3784

p-hydroxylation was shown to occur at some stage beyond the reduction of cytochrome P450. Experimental values in Ref. 11 are given with two decimals, and therefore four decimals should be taken into account after logarithmic transformation.

The set of weighted paths was truncated at l = 4, because paths P_5, P_6 and P_7 are present only in a few of the alcohols studied. We considered models with two, three and four descriptors. To measure the model quality, we used fit statistical parameters: the correlation coefficient (R) and standard error of estimate (S) and the corresponding leave-one-out cross-validated parameters (R_{cv}, S_{cv}). Standard errors of estimate between the experimental and fitted (S) and leave-one-out cross-validated (S_{cv}) values were calculated using the number of compounds in the denominator, in the same way as it was done in our earlier works.[16,17] We searched for the optimum value of weight x by varying it in the range 0.00–5.00, depending on the number of

descriptors used in the modeling. The standard error of estimate S and leave-one-out cross-validated error of estimate S_{cv} were used as criteria for the selection of the optimum value of x.

Best Models with Two Descriptors

Six models with two descriptors are possible out of the set of four descriptors. For each value of x from the range 0.00–5.00, all the best 2-descriptor models were always based on P_1 and P_3 weighted paths. In Table III, we give the standard errors of estimate S and leave-one-out cross-validated errors of estimate S_{cv} for various values of x.

TABLE III

The change of the standard error of estimate S and leave-one-out cross-validated error of estimate S_{cv} for various values of x when the weighted P_1 and P_3 paths are used as descriptors. We varied x from 0.00 to 5.00, but only the results in the limited range from 1.20 to 2.30 are given. The minimum values of S and S_{cv} are denoted by asterisks.

x	S	S_{cv}
1.20	0.2667	0.3346
1.30	0.2608	0.3286
1.40	0.2566	0.3249
1.50	0.2536	0.3226
1.60	0.2516	0.3215
1.70	0.2504	*0.3210
1.80	0.2496	0.3211
1.90	0.2493	0.3216
2.00	*0.2492	0.3223
2.05	0.2493	0.3228
2.10	0.2494	0.3232
2.15	0.2495	0.3237
2.20	0.2497	0.3242
2.25	0.2499	0.3248
2.30	0.2501	0.3253

It follows from Table III that the minimum value of S is obtained at x = 2.00, while S_{cv} reaches minimum at x = 1.70. Below, we explicitly give models for x = 1.70, 1.80, 1.90 and 2.00:

(i) $x = 1.70$

$$\log(1/I_{50}) = -2.389\ (\pm 0.341) + 0.620\ (\pm 0.088)\ P_1 - 0.321\ (\pm 0.048)\ P_3 \quad (1)$$

$$N = 19 \quad R = 0.874 \quad R_{cv} = 0.785 \quad S = 0.2504 \quad S_{cv} = 0.3210$$

(ii) $x = 1.80$

$$\log(1/I_{50}) = -2.395\ (\pm 0.343) + 0.605\ (\pm 0.085)\ P_1 - 0.303\ (\pm 0.045)\ P_3 \quad (2)$$

$$N = 19 \quad R = 0.874 \quad R_{cv} = 0.785 \quad S = 0.2496 \quad S_{cv} = 0.3211$$

(iii) $x = 1.90$

$$\log(1/I_{50}) = -2.403\ (\pm 0.346) + 0.590\ (\pm 0.083)\ P_1 - 0.287\ (\pm 0.043)\ P_3 \quad (3)$$

$$N = 19 \quad R = 0.875 \quad R_{cv} = 0.784 \quad S = 0.2493 \quad S_{cv} = 0.3216$$

(iv) $x = 2.00$

$$\log(1/I_{50}) = -2.412\ (\pm 0.349) + 0.577\ (\pm 0.082)\ P_1 - 0.274\ (\pm 0.041)\ P_3 \quad (4)$$

$$N = 19 \quad R = 0.875 \quad R_{cv} = 0.783 \quad S = 0.2492 \quad S_{cv} = 0.3223$$

There is little difference between these four models. However, if R and S are the criteria of model quality, then model (4) is preferred. On the other hand, if R_{cv} and S_{cv} are taken to be quality criteria, then model (1) is slighly better than any of the other three models given above.

Best Models with Three Descriptors

Three models with three descriptors are possible out of the set of four descriptors. For each x value from the range 0.00–5.00, the best models obtained were based on P_1, P_2 and P_3 weighted paths. In Table IV, we give the standard errors of estimate S and leave-one-out cross-validated errors of estimate S_{cv} for various values of x.

TABLE IV

The change of the standard error of estimate S and leave-one-out cross-validated error of estimate S_{cv} for various values of x when the weighted P_1, P_2 and P_3 paths are used as descriptors. We varied x from 0.00 to 5.00, only the results in the limited range from 0.70 to 1.70 are given. The minimum values of S and S_{cv} are denoted by asterisks.

x	S	S_{cv}
0.70	0.1986	0.2620
0.80	0.1911	0.2538
0.90	0.1860	0.2490
1.00	0.1827	0.2465
1.10	0.1809	*0.2457
1.20	0.1801	0.2458
1.30	*0.1800	0.2468
1.40	0.1804	0.2482
1.50	0.1812	0.2500
1.60	0.1822	0.2520
1.70	0.1834	0.2542

It follows from Table IV that the minimum value of S is obtained at x = 1.30, while S_{cv} reaches minimum at x = 1.10. Below, we explicitly give models for x = 1.10, 1.20 and 1.30:

(i) $x = 1.10$

$$\log(1/I_{50}) = -2.237\ (\pm 0.250) + 0.894\ (\pm 0.087)\ P_1 - 0.287\ (\pm 0.065)\ P_2 - 0.316\ (\pm 0.065)\ P_3 \quad (5)$$

$$N = 19 \quad R = 0.936 \quad R_{cv} = 0.880 \quad S = 0.1809 \quad S_{cv} = 0.2457$$

(ii) $x = 1.20$

$$\log(1/I_{50}) = -2.235\ (\pm 0.250) + 0.863\ (\pm 0.083)\ P_1 - 0.272\ (\pm 0.064)\ P_2 - 0.296\ (\pm 0.060)\ P_3 \quad (6)$$

$$N = 19 \quad R = 0.937 \quad R_{cv} = 0.880 \quad S = 0.1801 \quad S_{cv} = 0.2458$$

TABLE V

The change of the standard error of estimate S and leave-one-out cross-validated error of estimate S_{cv} for various values of x when the weighted P_1, P_2, P_3 and P_4 paths are used as descriptors. We varied x from 0.00 to 5.00, but only the results in the limited range from 0.70 to 1.70 are given. The minimum values of S and S_{cv} are denoted by asterisks.

x	S	S_{cv}
0.70	0.1371	0.2054
0.80	0.1263	0.1952
0.90	0.1191	0.1864
1.00	0.1153	0.1797
1.10	*0.1143	0.1757
1.20	0.1156	*0.1749
1.30	0.1186	0.1768
1.40	0.1226	0.1811
1.50	0.1272	0.1870
1.60	0.1321	0.1940
1.70	0.1371	0.2014

(iii) $x = 1.30$

$$\log(1/I_{50}) = -2.235\ (\pm 0.250) + 0.836\ (\pm 0.081)\ P_1 - 0.258\ (\pm 0.064)\ P_2 - 0.279\ (\pm 0.057)\ P_3 \quad (7)$$

$$N = 19 \quad R = 0.937 \quad R_{cv} = 0.879 \quad S = 0.1800 \quad S_{cv} = 0.2468$$

There is little difference between these three models. However, if R and S are the criteria of model quality, then model (7) is preferred. On the other hand, if R_{cv} and S_{cv} are taken to be quality criteria, then model (5) is slighly better than any of the other three models given above. However, models (5)–(7) possess better statistical parameters than models (1)–(4).

Best Model with Four Descriptors

In Table V, we give the standard error of estimate S and leave-one-out cross-validated error of estimate S_{cv} for various values of x.

It follows from Table V that the minimum value of S is obtained at x = 1.10, while S_{cv} reaches minimum at x = 1.20. Below, we explicitly give both models:

(i) $x = 1.10$

$$\log(1/I_{50}) = -3.164\ (\pm 0.259) + 1.280\ (\pm 0.102)\ P_1 - 0.412\ (\pm 0.050)\ P_2 - 0.305\ (\pm 0.042)\ P_3 - 0.257\ (\pm 0.056)\ P_4 \quad (8)$$

$$N = 19 \quad R = 0.975 \quad R_{cv} = 0.943 \quad S = 0.1143 \quad S_{cv} = 0.1757$$

(ii) $x = 1.20$

$$\log(1/I_{50}) = -3.185\ (\pm 0.270) + 1.249\ (\pm 0.103)\ P_1 - 0.389\ (\pm 0.050)\ P_2 - 0.297\ (\pm 0.040)\ P_3 - 0.245\ (\pm 0.055)\ P_4 \quad (9)$$

$$N = 19 \quad R = 0.974 \quad R_{cv} = 0.943 \quad S = 0.1156 \quad S_{cv} = 0.1749$$

There is little difference between these two models and both models are superior to all the other models reported above.

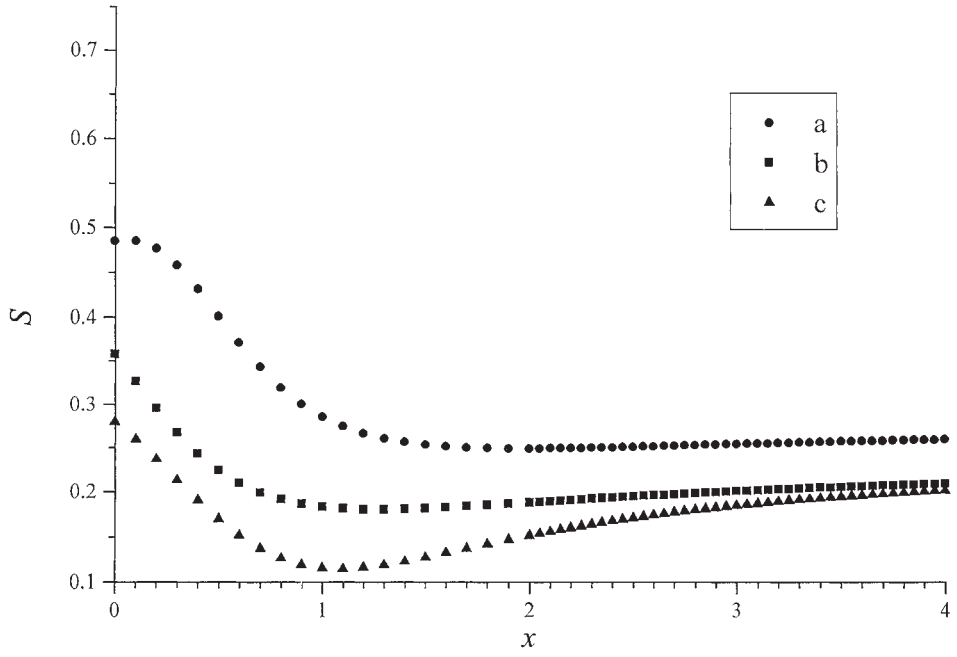


Figure 2. Scatter plot between fit standard errors (S) and x values in the range 0.00–4.00 for the models containing two (P_1, P_3 – filled circles; a), three (P_1, P_2, P_3 – filled squares; b), and four (P_1, P_2, P_3, P_4 – filled triangles; c) weighted paths as descriptors.

The Preferred Model

All the values of S and S_{cv} parameters for the best models containing two, three or four path numbers (descriptors) in the range of x between 0.00 and 4.00 are given in Figure 2 (for S) and Figure 3 (for S_{cv}).

It is important to compare the S and S_{cv} values of the two-, three- and four-descriptor models for x = 1.00 (without optimization of x) with the S and S_{cv} values of the best models (having minimal S or S_{cv}). One can see that, in each case, all the best models are obtained for weighted paths calculated for x values between 1.10 and 2.00. In order to select the best one among the models developed, we have to take into account experimental errors of 17 experimental I_{50} values (Cohen and Mannering[11] did not give experimental errors for 3-hexanol and 2-methyl-1-propanol). To do that, we have to transform the calculated (fitted and cross-validated) $\log(1/I_{50})$ obtained by the best models from Tables III–V from the logarithmic scale to the linear scale. The lowest mean absolute error for 19 fitted values obtained for the model containing four weighted paths P_1, P_2, P_3, and P_4 (calculated

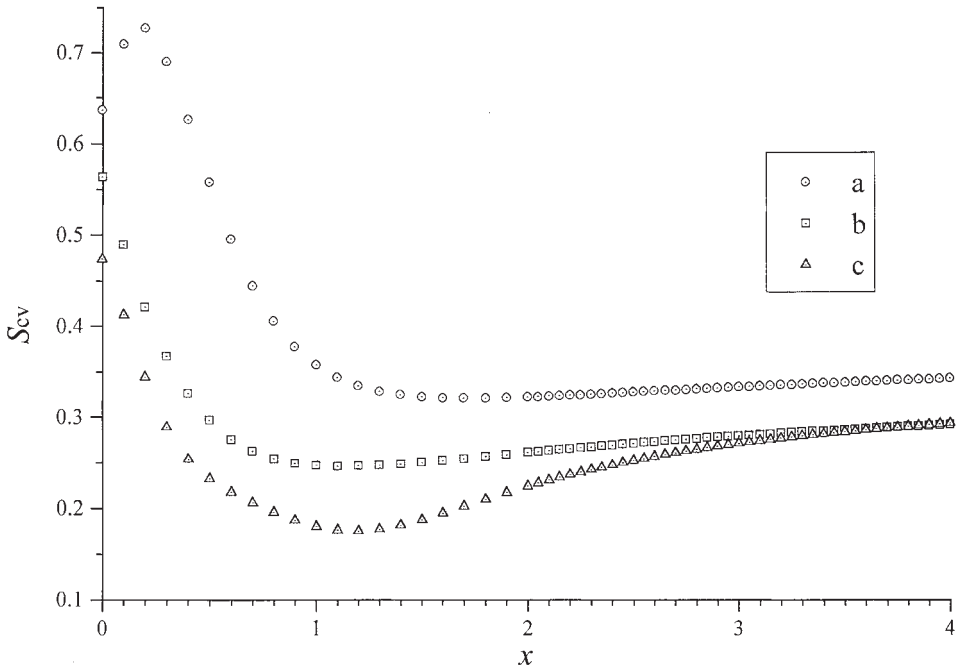


Figure 3. Scatter plot between cross-validated standard errors (S_{cv}) and x values in the range 0.00–4.00 for the models containing two (P_1, P_3 – circles; a), three (P_1, P_2, P_3 – squares; b), and four (P_1, P_2, P_3, P_4 – triangles; c) weighted paths as descriptors.

with $x = 1.20$) is 0.87, and it is comparable to the mean absolute error (0.66) for 17 experimetal values given by Cohen and Mannering.[11] Also, this model has the lowest cross-validated mean absolute error.

From Figure 3 one can see that the two-descriptor models for the x-values in the range 0.00–4.00 are not stable (measured by cross-validated parameters S_{cv} and R_{cv}), *i.e.* cross-validated parameters are far from the corresponding fitted values.

In Figures 4 and 5 we give plots between the experimental and calculated values of $\log(1/I_{50})$ for fit and cross-validated models with four descriptors.

Comparison with Models from the Literature

Cohen and Mannering[11] generated three models:

(i) Linear model using the logarithm of the partition coefficient ($\log P$):

$$\log(1/I_{50}) = 1.83\ (\pm 0.44) + 0.82\ (\pm 0.33)\ \log P \qquad (10)$$

$$N = 17 \quad R = 0.80 \quad S = 0.53$$

(ii) Quadratic model with $\log P$:

$$\log(1/I_{50}) = 1.75\ (\pm 0.38) + 1.50\ (\pm 0.62)\ \log P - 0.36\ (\pm 0.29)\ (\log P)^2 \qquad (11)$$

$$N = 17 \quad R = 0.98 \quad S = 0.44$$

(iii) Multivariate model with $\log P$ and Taft's steric substituent (E_s):

$$\log(1/I_{50}) = 2.03\ (\pm 0.29) + 2.09\ (\pm 0.50)\ \log P - 0.58\ (\pm 0.22)\ (\log P)^2 + 0.63(\pm\ 0.31)\ E_s \qquad (12)$$

$$N = 17 \quad R = 0.95 \quad S = 0.29$$

Judging by the S values, our models are better than Cohen and Mannering's models, but their models were developed for a reduced set of 17 compounds, and they also included methanol. For this reason, these models can

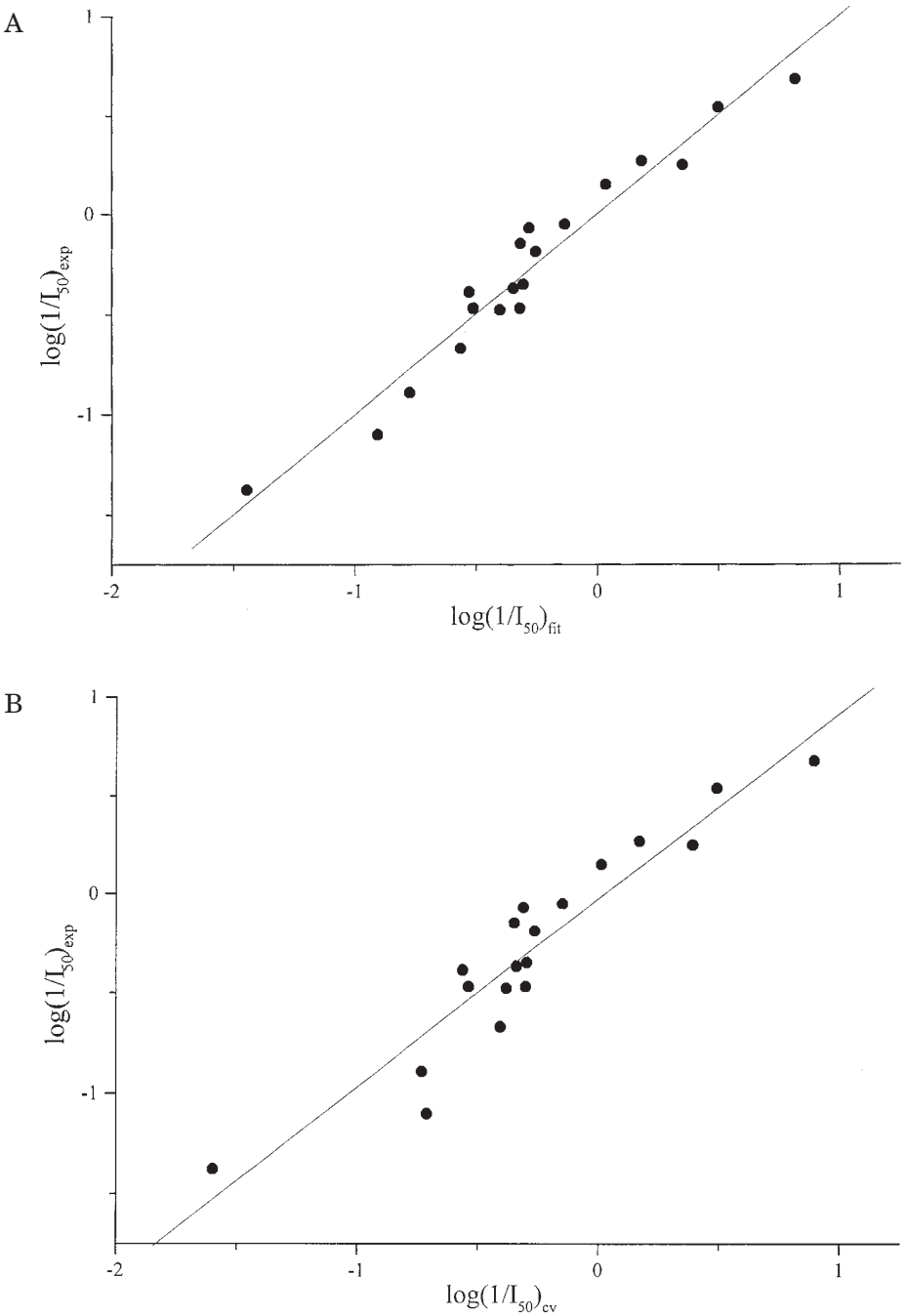


Figure 4. Scatter plots between the experimental and calculated values of $\log(1/I_{50})$ for fit (A) and cross-validated (B) models with four weighted path numbers.

not be strictly compared to the models developed in this study. In addition, statistical parameters of our models are superior to those produced by Basak and Gute[18] using the atom pair (the best model: R = 0.878; We do not give S values given in Ref. 18 because these values were not correctly calculated) and Euclidean distance (the best model: R = 0.707) similarity approaches.

CONCLUDING REMARK

Among the available QSAR models for predicting inhibition of microsomal *p*-hydroxylation of aniline by alipahtic alcohols, the models that are based on the weighted path numbers possess very good statistical parameters. Thus, the (modified) idea of Platt[19] that path numbers might be useful molecular descriptors, suggested more than fifty years ago, is finally bearing fruits.[2–4,10,15,20]

Acknowledgements. – This work was supported in part by Grants 079301 (D. A.), 00980606 (B. L., S. N., N. T.) and 06MN407 (B. L.) awarded by the Ministry of Science and Technology of the Republic of Croatia.

REFERENCES

1. QSAR is the standard acronym for quantitative structure-activity relationships.
2. M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci*. **39** (1999) 261–266.
3. M. Randić and M. Pompe, *SAR QSAR Environ. Res.* **10** (1999) 451–471.
4. M. Randić and S. C. Basak, *SAR QSAR Environ. Res*. **11** (2000) 1–23.
5. B. Lučić, S. Nikolić, N. Trinajstić, A. Jurić, and Z. Mihalić, *Croat. Chem. Acta* **68** (1995) 417–434.
6. S. Nikolić and N. Trinajstić, *SAR QSAR Environ. Res.* **9** (1998) 117–126.
7. S. Nikolić, N. Trinajstić, D. Amić, D. Bešlo, and S. C. Basak, in: M. V. Diudea (Ed.), *QSPR/QSAR Studies by Molecular Descriptors*, Nova Science Publishers, Hungtington, 2000, pp. 71–89.
8. L. H. Hall and L. B. Kier, *Bull. Environ. Contam. Toxicol.* **32** (1984) 354–362.
9. S. H. Roth, *Fed. Proc.* **39** (1980) 1595–1599.
10. D. Amić, S. C. Basak, B. Lučić, S. Nikolić, and N. Trinajstić, Structure-Water Solubility Modeling of Aliphatic Alcohols Using the Weighted Path Numbers, reported in part at *QSAR 2000 – Crossroads to the XXI Century – The Ninth International Workshop on Quantitative Structure-Activity Relationships in Environmental Sciences* (Bourgas, Bulgaria, September 16–20, 2000) and submitted to *SAR QSAR Environ. Res*.
11. G. M. Cohen and G. J. Mannering, *Mol. Pharmacol*. **9** (1973) 383–397.
12. N. Trinajstić, *Chemical Graph Theory*, 2nd revised edition, CRC Press, Boca Raton, FL, 1992, pp. 35–37.
13. W. T. Tutte, *Connectivity in Graphs*, University of Toronto Press and Oxford University Press, London, 1966, p. 28.

14. M. Randić, *Croat. Chem. Acta* **64** (1991) 43–54.
15. M. Randić and S. C. Basak, in: D. K. Sinha, S. C. Basak, R. K. Mohanty, and I. N. Busamallick (Eds.), *Some Aspects of Mathematical Chemistry*, Visva-Bharati University Press, Santiniketan, India, in press.
16. B. Lučić, D. Amić, and N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **40** (2000) 403–413.
17. B. Lučić, I. Lukovits, S. Nikolić, and N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **41** (2001), in press.
18. S. C. Basak and B. D. Gute, in: B. L. Johnson, S. Xintaras, and J. S. Andrews, Jr. (Eds.), *Hazardous Waste – Impacts on Human and Ecological Health*, Princeton Sci. Publ. Co., Princeton, NJ, 1997, pp. 492–504. In their work the Basak and Gute used log values of experimental data provided by Cohen and Mannering (Ref. 11).
19. J. Platt, *J. Chem. Phys.* **15** (1947) 419–420.
20. A. T. Balaban, C. Catana, M. Dawson, and I. Niculescu-Duvaz, *Rev. Roum. Chim*. **35** (1990) 997–1003.

SAŽETAK

Predviđanje inhibicije mikrosomalne *p*-hidroksilacije anilina alifatskim alkoholima: QSAR postupak temeljen na vaganim staznim brojevima

Dragan Amić, Bono Lučić, Sonja Nikolić i Nenad Trinajstić

Upotrebljeni su vagani stazni brojevi za gradnju QSAR modela za pretkazivanje inhibicije mikrosomske *p*-hidroksilacije anilina alifatskim alkoholima. Razmatrani su modeli s dva, tri i četiri vagana stazna broja. Kvaliteta modela je određivana pomoću ugođenih i unakrsno vrednovanih statističkih parametara. Najbolje statističke parametre posjeduju modeli s četiri vagana stazna broja. Usporedba s modelima iz literature daje prednost modelima s vaganim staznim brojevima.