Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry Letters 15 (2005) 421–425
Bioorganic & Medicinal Chemistry Letters

Novel estimation of lipophilicity using ^{13}C NMR chemical shifts as molecular descriptor

Padmakar V. Khadikar,[a,*] Vimukta Sharma[b] and R. G. Varma[c]

[a]*Research Division, Laxmi Fumigation and Pest Control Pvt Ltd, 3, Khatipura, Indore 452 007, India*
[b]*Department of Pharmacy, Maharaja Ranjeetsingh College, Indore, India*
[c]*Department of Chemistry, PMB Gujarati Science College, Indore, India*

Received 8 July 2004; accepted 21 October 2004

Abstract—This paper describes the use of ^{13}C NMR chemical shift as molecular descriptor (molecular parameter) for modeling lipophilicity ($\log P$). A set of 32 alcohols were chosen for this purpose. The regression analysis of the data showed that ^{13}C NMR chemical shifts of these alcohols can be used as a molecular descriptor (molecular property) for modeling the lipophilicity ($\log P$). Better results are obtained by introducing an indicator parameter.

1. Introduction

Balasubramanian[1] has reviewed applications of combinatory and graph theory to spectroscopy. In the area of NMR spectroscopy these theories are very useful. Randic[2,3] has shown that the graph theoretical techniques could also be used to obtain the chemical shifts of the nuclei. They have developed[4] a computer code for listing equivalent classes of graphs, which works for most of the small, not transitive, and non-isospectral graphs. Furthermore, Duvenbeek[5] has discussed topological and geometral approaches to develop models for the prediction of ^{13}C NMR shifts and used $\sum$ ^{13}C NMR chemical shifts ($\sum$ $^{13}C_n$) as a molecular property.

It is well known that the presence of attached and nearby carbon atoms have a profound effect on ^{13}C NMR chemical shift. That is, topology of the molecule plays a dominant role in estimating ^{13}C NMR chemical shift.

Conversely, we can say that likewise ^{13}C NMR shifts also accounts for the topology of the molecules.

In our earlier communication,[6] while predicting ^{13}C NMR chemical shifts for 2,6-, and 2,7-disubstituted naphthalene, we have shown that $\sum$ ^{13}C can be used as a molecular property, which in turn can be modeled by both Wiener (W)[7] and Szeged (Sz)[8,9] indices. In still another communication[10] we have used PI (Padmakar–Ivan) index for predicting ^{13}C NMR chemical shifts of alkanes and cycloalkanes as well as their coupling constant.[11]

At this stage it is interesting to mention that although individual NMR chemical shifts for different atoms have received wide attention, it is somewhat surprising that there is hardly any study devoted to the collection of NMR chemical shifts for the purpose of drug modeling. It was shown that the average ^{13}C NMR chemical shifts of alkenes display regularity in isomeric variations. Such regulations are analogous to the isomeric variations of numerous thermodynamic properties of alkenes. Our results[6,10,11] also show that individual NMR chemical shifts can also be modeled topologically.

The aforementioned results prompted us that ^{13}C NMR chemical shift can be considered as a molecular property and that it can be used as a molecular descriptor for modeling physicochemical properties and biological activity of organic molecules acting as drugs. The results as discussed below show that this is found to be the case in the present study also.

Keywords: Chemical shift; ^{13}C NMR; Regression analysis; Molecular descriptor; QSAR.
* Corresponding author. Tel.: +91 731 2531906; fax: +91 731 2763618; e-mail: pvkhadikar@rediffmail.com

0960-894X/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2004.10.075

2. Results and discussion

The set of 32 alcohols, their ^{13}C NMR chemical shift, lipophilicity ($\log P$) and the assumed indicator parameters are given in Table 1. The details regarding this are given in the experimental section of this paper.

Application of regression analysis[12] showed that out of this set of 32 compounds, **2**, **6**, and **13** are outliers. They are, therefore, deleted from further regression analysis. At present we cannot provide convincing proof for their deletion and can consider it to be due to the regression procedure.

A comparison of correlation matrices for the set of 32 compounds (Table 2) and that for 29 compounds, after deleting compounds **2**, **6**, and **13** (Table 3), indicates that correlatedness amongst lipophilicity ($\log P$), ^{13}C NMR and indicator parameters (IP_1, IP_2, IP_3) is considerably improved for the new set of 29 compounds. Hence, the following results and discussion pertain to this set of 29 compounds. It is worthy to comment on the indicator parameters used. These are dummy parameters to account for those structural features not covered under the molecular descriptor used. They have only two values, 1 (when that structural feature is present in the structure) or 0 (when that structural feature is absent in the molecule). We have used $IP_1 = 1$ for the presence of a primary alcoholic group, in the absence of which $IP_1 = 0$. Similarly, IP_2 and IP_3 are taken as 1, for the presence of a secondary alcoholic group and a methyl group, respectively. Once again, in the absence of such structural features, both the indicator parameters IP_2 and IP_3 are 0 (zero).

A perusal of Table 3 shows that even in mono-parametric regression ^{13}C NMR chemical shift will be a good molecular descriptor for modeling lipophilicity ($\log P$) of the set of alcohols used. Also, that out of the three indicator parameters, the indicator parameter IP_1 will be more useful in obtaining multiparametric model(s). In view of this we have adopted maximum R^2-method[12] and carried out step-wise regression analysis. We have four correlating parameters: ^{13}C, IP_1, IP_2, and IP_3 and looking to the sample size and in accordance with the 'Rule of Thumb', we can go up to four-parametric (tetra-parametric) regression analysis. The results of simple regression, as well as those of step-wise regressions starting from mono-parametric up to tetra-parameteric regressions are given in Table 4.

In accordance with simple (mono-parametric) regression, the lipophilicity ($\log P$) can be modeled according to the following regression equation:

$$\log P = -3.9700 + 0.0774\ (\pm 0.0021)^{13}C$$
$$n = 29, \quad \mathrm{Se} = 0.5287, \quad R(r) = 0.7576,$$
$$F = 36.72, \quad Q = 1.4239 \qquad (1)$$

Table 1. Alcohols, their $\log P$, ^{13}C values, and indicator parameters (IP_1, IP_2, IP_3)

Compd no.	Name	$\log P$	^{13}C	I_1	I_2	I_3
1	Methanol	−0.764	49.0	1	1	0
2	Ethanol	−0.235	57.0	1	1	0
3	Propanol	0.294	63.6	1	1	0
4	Butanol	0.823	61.4	1	1	0
5	Pentanol	1.352	61.8	1	1	0
6	Hexanol	1.881	61.9	1	1	0
7	Isopropanol	0.154	63.4	1	1	0
8	2-Butanol	0.603	68.7	0	1	0
9	2-Pentanol	1.132	67.0	0	1	0
10	2-Hexanol	1.661	67.2	0	1	0
11	3-Pentanol	1.132	73.8	0	1	0
12	3-Hexanol	1.661	72.3	0	1	0
13	3-Heptanol	2.190	72.6	0	1	0
14	4-Heptanol	2.190	70.6	0	1	0
15	4-Octanol	2.680	70.9	0	1	0
16	5-Nonanol	1.572	71.1	0	0	0
17	Isobutanol	0.805	68.9	1	0	0
18	Tetra-butanol	0.532	68.4	1	0	0
19	Neopentanol	1.664	72.6	1	0	0
20	2-Methyl-pentanol	0.693	66.9	0	0	1
21	3-Methyl-butanol	1.280	60.2	0	0	1
22	3-Methyl-2-butanol	1.280	72.0	0	0	1
23	4-Methyl-2-butanol	1.687	65.2	0	0	1
24	4-Methyl-3-pentanol	1.687	77.3	0	0	1
25	3,3-Dimethyl-butanol	1.808	58.9	0	0	1
26	2,3-Dimethyl-2-butanol	1.529	72.2	0	0	1
27	3,3-Dimethyl-2-butanol	1.480	74.8	0	0	1
28	4,4-Dimethyl-3-butanol	2.154	80.9	0	0	1
29	2,4-Dimethyl-3-pentanol	2.148	80.4	0	0	1
30	2,3,3-Trimethyl-2-butanol	1.996	74.1	0	0	1
31	2,4,4-Trimethyl-3-pentanol	2.615	82.8	0	0	1
32	2,2,4,4-Tetramethyl-3-pentanol	3.082	84.7	0	0	1

Table 2. Correlation matrix for all the set of 32 compounds

	$\log P$	^{13}C	IP_1	IP_2	IP_3
$\log P$	1.0000				
^{13}C	0.4708	1.0000			
IP_1	−0.6070	−0.3623	1.0000		
IP_2	0.1907	0.0306	−0.4217	1.0000	
IP_3	0.3983	0.3139	−0.5577	−0.5174	1.0000

Table 3. Correlation matrix for the set of 29 compounds, that is, after deleting compounds **2**, **6**, and **13**

	$\log P$	^{13}C	IP_1	IP_2	IP_3
$\log P$	1.0000				
^{13}C	0.7576	1.0000			
IP_1	−0.6238	−0.4963	1.0000		
IP_2	0.1030	0.0540	−0.3810	1.0000	
IP_3	0.4681	0.3950	−0.5564	−0.5564	1.0000

Table 4. Regression parameters and quality of correlation for the proposed models for 29 compounds

Model	Parameter used	Se	$R(r)$	R^2_A	F	Q	6PE
1	^{13}C	0.5287	0.7576	–	36.372	1.4330	0.1584
2	^{13}C, IP_1	0.4845	0.8096	0.6289	24.729	1.6710	0.0740
3	^{13}C, IP_2	0.5363	0.7601	0.5453	17.792	1.4192	0.1782
4	^{13}C, IP_3	0.5174	0.7791	0.5768	20.082	1.5058	0.0810
5	^{13}C, IP_1, IP_2	0.4915	0.8118	0.6181	16.104	1.6517	0.4980
6	^{13}C, IP_1, IP_3	0.4915	0.8118	0.6181	16.104	1.6517	0.4980
7	^{13}C, IP_2, IP_3	0.4915	0.8118	0.6181	16.104	1.6517	0.4980
8	^{13}C, IP_1, IP_2, IP_3	0.4915	0.8118	0.6181	16.104	1.6517	0.4980

Here and thereafter n is the number of compounds, Se is the standard error of estimation, r is the simple correlation coefficient, R is the multiple correlation coefficient, F is the Fisher's statistic and Q is the quality factor,[13,14] $Q = R(r)/\text{Se}$.

The positive coefficient of ^{13}C NMR chemical shift in the above Eq. 1 indicates that the lipophilicity ($\log P$) is directly (linearly) related to the magnitude of ^{13}C NMR chemical shift.

The step-wise regression analysis has indicated that there are only three possible bi-parametric regression models, which can yield better results than the mono-parameteric regression discussed above. The statistics of each of these three bi-parametric models are presented in Table 4. The perusal of Table 4 shows that a bi-parametric model containing ^{13}C and IP_2 (or IP_3) yielded slightly better results than the simple regression model. However, the bi-parametric model containing ^{13}C and IP_1 has quite improved statistics. This model is found as below:

$$\log P = -0.0609\ (\pm 0.0136)^{13}\text{C} - 0.5750\ (\pm 0.2319)\text{IP}_1 - 2.6933$$
$$n = 29, \quad \text{Se} = 0.4845, \quad R = 0.8096,$$
$$R^2_A = 0.6289, \quad F = 24729, \quad Q = 1.6710 \qquad (2)$$

Once again, the positive coefficient of ^{13}C indicates its favorable contribution for modeling, monitoring, and estimating $\log P$. However, the coefficient of the indicator parameter IP_1 is negative meaning, thereby, that the presence of $-CH_2OH$ (primary alcoholic group) has a negative effect on the exhibition of $\log P$.

In an attempt to obtain an even better model we have carried out tri-parametric regression analysis. All the three tri-parametric models (Table 4) gave similar results. However, the physical significance and the biological relevance of each of these models are quite different, which can be judged by the following regression expressions:

$$\log P = -0.0597\ (\pm 0.0140)^{13}\text{C} - 0.6288\ (\pm 0.2577)\text{IP}_1 - 0.1144\ (\pm 0.2241)\text{IP}_2 - 2.5630$$
$$n = 29, \quad \text{Se} = 0.4915, \quad R = 0.8118,$$
$$R^2_A = 0.6181, \quad F = 16.104, \quad Q = 1.6517 \qquad (3)$$

$$\log P = -0.0597\ (\pm 0.0140)^{13}\text{C} - 0.5143\ (\pm 0.2635)\text{IP}_1 - 0.1144\ (\pm 0.2241)\text{IP}_3 - 2.6775$$
$$n = 29, \quad Se = 0.4915, \quad R = 0.8118,$$
$$R^2_A = 0.6181, \quad F = 16.104, \quad Q = 1.6517 \qquad (4)$$

$$\log P = -0.0597\ (\pm 0.0140)^{13}\text{C} - 0.5143\ (\pm 0.2635)\text{IP}_2 - 0.6288\ (\pm 0.2577)\text{IP}_3 - 3.1919$$
$$n = 29, \quad \text{Se} = 0.4915, \quad R = 0.8118,$$
$$R^2_A = 0.6181, \quad F = 16.104, \quad Q = 1.6517 \qquad (5)$$

Out of these three tri-parametric models, the models expressed by Eqs. 3 and 4 are discarded on the ground that in Eq. 3 the coefficient of IP_2 term is very smaller than its standard division. Similarly, in Eq. 4 the coefficient of IP_3 term is likewise smaller than its standard division. Such models are not allowed statistically. In contrast to these models, the model based on ^{13}C, IP_2, and IP_3 is statistically allowed and is the only tri-parametric model found better than the bi-parametric model discussed above. It is interesting to mention that this model (Eq. 5) has a positive coefficient for all the three parameters involved. Also, that the coefficient of all the three correlation parameters are higher than that corresponding standard deviations. This model once again suggests that the value of $\log P$ is directly related to the magnitude of ^{13}C NMR chemical shift. The positive coefficients of IP_2 and IP_3 terms in Eq. 5 indicates that a secondary alcoholic group and methyl substitution are favorable for the exhibition of $\log P$.

Our attempts for obtaining an even better model failed as the only possible tetra-parametric model is the one based on the combination of ^{13}C, IP_1, IP_2, and IP_3. However, it yielded similar statistics as that for tri-parametric models discussed above. Also, in this model the coefficients of IP_2 and IP_3 were quite smaller than their standard deviations. Furthermore, there is no change in R^2_A value, it remained the same ($R^2_A = 0.6181$) indicating that the added parameter does not have a favorable contribution for the exhibition of $\log P$.

In order to confirm our results we have estimated $\log P$ values from the best tri-parametric model and compared them with the observed values of $\log P$. Such a comparison is shown in Table 5. Also, a standardised residual plot has indicated that there is no outlier in this model.

The models obtained and recorded in Table 4 need further investigation on the basis of R^2_A (adjusted R^2). Usually R^2 increases with an increase in the correlating parameters, however, R^2_A decreases if the added parameter does not have a significant contribution to the model. In our case, R^2_A for the bi-parametric model containing ^{13}C and IP_1 as the correlating parameters has a value of $R^2_A = 0.6289$. However, all the three tri-parametric models have $R^2_A = 0.6181$, a value smaller than the bi-parametric model. This means that the added parameter to the previous bi-parametric model containing ^{13}C and IP_1 does not have a favorable contribution to the deployed tri-parametric models. However, the value of R increased slightly from 0.8096 to 0.8118.

Table 5. Found and estimated $\log P$ using the models 2 and 5

Comp no.	$\log P$ (Obs)	Estimated $\log P$			
		Model 2		Model 5	
		Est.	Res.	Est.	Res.
1	−0.764	−0.289	−0.475	−0.265	−0.499
2	−0.235	Outlier	—	Outlier	—
3	0.294	0.606	0.018	0.606	−0.312
4	0.823	0.461	0.121	0.464	0.359
5	1.352	0.497	0.855	0.499	0.853
6	1.882	Outlier	—	Outlier	—
7	0.154	0.594	0.440	0.594	−0.440
8	0.683	1.492	−0.809	1.425	−0.742
9	1.132	1.389	−0.257	1.324	−0.192
10	1.661	1.401	0.260	1.336	0.325
11	1.132	1.803	−0.671	1.730	−0.598
12	1.661	1.711	−0.050	1.640	0.021
13	2.190	Outlier	—	Outlier	—
14	2.190	1.730	0.460	1.658	0.532
15	2.680	1.608	1.072	1.539	1.141
16	1.572	1.638	−0.066	1.569	0.003
17	0.805	0.929	−0.124	0.923	−0.118
18	0.532	0.899	−0.367	0.893	−0.361
19	1.662	1.155	−0.507	1.144	0.518
20	0.693	1.382	−0.689	1.432	−0.379
21	1.280	0.974	0.306	1.032	0.248
22	1.280	1.693	−0.413	1.737	−0.457
23	1.687	1.279	0.408	1.331	0.356
24	1.687	2.016	−0.329	2.053	−0.366
25	1.808	0.895	0.913	0.955	0.853
26	1.529	1.705	−0.176	1.749	−0.220
27	1.480	1.864	−0.384	1.904	−0.424
28	2.154	2.235	−0.081	2.268	−0.114
29	2.148	2.205	−0.057	2.239	−0.091
30	1.996	1.821	0.175	1.862	0.134
31	2.615	2.351	0.264	2.382	0.233
32	3.082	2.467	0.615	2.495	0.587

This means that though there is slight improvement in R value, all the three tri-parametric models are of less importance than the bi-parametric model containing ^{13}C and IP_1 as the correlating parameters. This is further confirmed by using a statistical parameter PE, as discussed below.

In order to judge the predictive power of the proposed models we have calculated Q values and presented them in Table 4, which shows that the mono-parametric model based on ^{13}C NMR shift alone has the least predictive power and that the tri-parametric as well as tetra-parametric models has similar predictive power. It also shows that the bi-parametric model using ^{13}C and IP_1 as the correlating parameters has the highest predictive power. A close look at Table 4 indicates that this bi-parametric model has more-or-less similar statistics. It is worthy to mention that under such a situation the model containing the lesser correlating parameter is considered the best model. Hence, the bi-parametric model containing ^{13}C and IP_1 is the most appropriate mode for modeling the lipophilicity of the compound used.

It is worth mentioning that by using the set of 32 alcohols chosen, the ^{13}C NMR shift proved to be useful NMR parameter as molecular descriptor for modeling $\log P$. The study is of importance but one may argue that it should be interesting to validate it differently: to try to predict $\log P$ of a compound that is not known, synthesize it and then determine $\log P$ experimentally. Also, one may argue without such validations it is rather difficult to imagine the validity of all these interesting studies. However, such arguments are of great value for the synthetic chemists involved in the preparation of series of compounds. The present study is based on well-known series of alcohols with their known ^{13}C NMR shifts, which we have used to propose its importance and usefulness as molecular descriptor.

For supporting our results we have calculated an interesting parameter called probable error of the coefficient of correlation (PE). This parameter is defined by the following expression:

$$PE = \frac{2}{3}\frac{1-r^2}{\sqrt{n}}$$

where, r is the coefficient of correlation and n is the number of compounds used. Based on the PE values the following recommendations are made: If,

(1) $r < PE$, r is not significant;
(2) $r > PE$, several times, at least three times greater, correlation is indicated, and
(3) $r > 6PE$, correlation is definitely good.

We have, therefore, calculated PE values for the proposed models and recorded them in Table 4 for comparison. We observed that all the proposed models have r values >6PE indicating all correlations attempted are definitely good. It is worth mentioning that 6PE for the bi-parametric model containing ^{13}C and IP_1 as the correcting parameter is the smallest (0.0740): in this case PE is several times larger than 6, and this is the most appropriate model for modeling lipophilicity of the set of compounds used.

3. Conclusions

From the results and discussion made above we concluded that ^{13}C NMR shift can be successfully used as a molecular descriptor for modeling, monitoring, and estimating lipophilicity ($\log P$) of the type of compounds used in the present study. A single parameter is found useful for a larger set of 29 compounds.

4. Experimental

Lipophilicity ($\log P$)—The lipophilicity ($\log P$) values were adopted from the previous work reported in the literature.

^{13}C *NMR shift*—The ^{13}C NMR chemical shifts were calculated from the corresponding NMR spectrum of the compound.

Regression analysis—All regression analysis were made using maximum R^2 method. The software provided by Istvan Lukovits was used for this purpose.

Acknowledgements

Authors thanks are due to Prof. Istvan Lukovits for providing the software for making regression analysis and to the referees for their valuable suggestions.

References and notes

1. Balasubramanian, K.. *Chem. Rev.* **1985**, *85*, 599.
2. Randic, H. *Int. J. Pharm. Chem.* **1983**, *23*, 1707.
3. Randic, H. *J. Math. Chem.* **1984**, *59*, 34.
4. Randic, H.; Trinajstic, N. *Theor. Chem. Acta* **1988**, *73*, 233.
5. Duvenbeek, C. *Topological and Geometrical Approaches to Develop Models for Prediction of ^{13}C NMR Shifts*; Bochum: GER, 1995.
6. Khadikar, P. V.; Pathre, S. V.; Shrivastava, A. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 2073.
7. Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
8. Gutman, I. *Graph Theory Notes*, New York, **1994**, *27*, 9.
9. Khadikar, P. V.; Deshpande, N. V.; Kale, P. P.; Dobrynin, A.; Gutman, I.; Domotor, G. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 547.
10. Khadikar, P. V.; Bajaj, A. V.; Mandloi, A. *Indian J. Chem.* **2002**, *41A*, 2067.
11. Khadikar, P. V.; Mandloi, M.; Bajaj, A. V. *Oxid. Commun.* **2004**, *27*, 23.
12. Chatterjee, S.; Hadi, A. S.; Price, B. *Regression Analysis by Examples*, 3rd ed.; Wiley: New York, 2000.
13. Pogliani, L. *Amino Acids* **1994**, *6*, 141.
14. Pogliani, L. *J. Phys. Chem.* **1996**, *100*, 18065.