A New Descriptor for Structure−Property and Structure−Activity Correlations†

Milan Randić*,‡,§ and Subhash C. Basak#

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311, Laboratory for Chemometrics, National Institute of Chemistry, Hajdrihova 19, Ljubljana, Slovenia, and Natural Resources Research Institute, University of Minnesota at Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received August 12, 2000

We consider an improvement of a multiple regression analysis (MRA) for the correlation of boiling points of alcohols using descriptors that involve a variable part. In the search for the best descriptors based on weighted paths we came upon a novel molecular descriptor, the use of which was apparently overlooked in the past. The novel descriptor counts paths of length three; however, only those associated with an oxygen atom are counted.

INTRODUCTION

"It is unusual for only one model to be compatible with experimental observations. Often data are not sufficiently extensive to discriminate among rival models, and new experiments must be designed to answer the outstanding questions. The statistical, graph theoretical, and sensitivity analysis methods ... can identify the areas for further investigation that are likely to produce significant new results."[1]

Structure−property and structure−activity studies continue to attract the attention of scientists of different backgrounds: organic chemists and biochemists, physical and theoretical chemists, theoretical physicists, mathematicians, computer scientists and statisticians, in particular, among others. It is clear that researchers of such diverse and different backgrounds are often more familiar with one or two of the diverse methodologies used. These include the following: quantum chemistry and ab initio computations; semiempirical approaches including molecular mechanics, structure coding, and computer graphics; synthesis planning; discrete mathematics; sensitivity analysis; etc. Statistical methodologies include pattern recognition, artificial neural networks, principal component analysis, partial least-squares approach, and, the oldest of all, multivariate regression analysis. An additional topic common to several of the methodologies to be mentioned is as follows: similarity/dissimilarity analysis, mining of large databases, and selection of molecular descriptors.

Hence, the QSAR and QSPR fields are in a way reminiscent of the building of the tower of Babylon. Although the goal appears common, people involved in the task not only speak different languages but also are not interested, not willing, or not capable of learning the "languages" of others. One of the misunderstandings accompanying such work is a lack of appreciation that models may be different, yet autonomous. By autonomous we mean that models have the same legitimacy even though they may employ a system of concepts that have no meaning outside a narrow area that concern such models. Such concepts, that include molecular descriptors, however, ought to have an interpretation *within* the model itself. The same concepts may have no simple meaning, or no meaning at all, when analyzed from the viewpoint of another model.[2]

The major models for structure−property-activity studies may be grouped into the four types (Table 1): empirical model; mathematical (calculus) model; mathematical (discrete) model; and graphical model (computer graphics). All these models have a common goal which is the representation and characterization of chemical structures. They differ in their input and their output information, that is, the way in which they represent chemical structures and what quantitative descriptors they extract from their model. The ultimate criterion for validation of any of the models is *the predictability*, that is the potential to characterize structures and properties that were not used in modeling. As we see from Table 1 each model has its own typical molecular descriptors, but a close view of such models will reveal that models may overlap to some extent. In Table 2 we listed additional nonobservables often met in chemistry merely to indicate that topological indices are no different from many of the concepts that most chemists are familiar with, like Kekulé valence structure, hybridization, potential curve, etc. Hence, HOMO−LUMO, a quantum chemical descriptor often used in QSAR, is as much an arbitrary descriptor as is the connectivity index or the Wiener index of chemical graph theory.[3] If, however, one thinks that HOMO−LUMO is not an arbitrary descriptor, because it has an apparent interpretation in quantum chemical models, then one has also to accept that the connectivity index and the Wiener index are not arbitrary descriptors, because they have (or ought to have) an interpretation in graph theoretical models of chemical structure. And indeed, they have, as has been recently discussed in a paper devoted to interpretation of topological indices.[4] To complain that topological indices have no physical meaning is equivalent to complaining that Hammett/Taft parameters have no topological meaning. Why should

* Corresponding author.

† Dedicated to the memory of Professor Robert S. Hansen (1918−1998), former Director of Ames Laboratory, Ames, IA, gracious supporter of Chemical Graph Theory in the time of need.

‡ Drake University.

§ National Institute of Chemistry.

University of Minnesota at Duluth.

10.1021/ci000116e CCC: $20.00
Published on Web 03/13/2001

Table 1. Four Major Computational Models for Study of Structure−Property-Activity Relationship

classical empirical models	mathematical models	mathematical models	computer models
	discrete mathematics	calculus	computer graphics
Hansch analysis	chemical graph theory	quantum chemistry	docking
linear free energy relation	Hückel MO	ab initio calculations	search for 3-D pharmacophore
molecular mechanics		statistical mechanics	
		classical trajectories	
molecular properties	topological indices	bond orders	
Hammett/Taft	indicator variables	HOMO−LUMO	
log P			

Table 2. Nonobservables of Classical Models, Graph Theoretical Models, and Quantum Chemical Models

classical/empirical	graph theory	quantum chemistry
bond dipole	connectivity index	bond orders
atomic charge	Wiener index	atom charge
potential curve	Hosoya Z index	resonance energy
van der Waals radii	Balaban J index	HOMO−LUMO
Kekulé valence structure	topological resonance energy	moments
Clar π-sextet	conjugated circuits	hybridization
	Hückel MO	Hückel MO
	Hückel ($4n+2$) rule	Hückel ($4n+2$) rule
molecular surface	Kekulé valence structure	atom polarizabilities
molecular volume	ID numbers	ring currents
pharmacophore	Clar π-sextet	configuration interaction
Periodic Table of Elements	Periodic Table of Isomers	molecular orbitals
		frontier orbitals
aromaticity	aromaticity	aromaticity

they have interpretations in models for which they have not been designed? One has to realize that all models have the same legitimacy; their relative utility will be judged by how well they can predict unavailable data.

QUESTIONS

Do we need additional descriptors although hundreds of molecular descriptors (graph theoretical as well as quantum chemical) are already available for use in structure−property and structure−activity studies? How can we establish whether the available molecular descriptors suffice for a *complete* characterization of molecules for QSPR and QSAR? When should we stop in improving a regression (and other data reduction schemes, like pattern recognition, principal component analysis, artificial neural networks, etc.)? How do we know that a particular regression equation cannot be further improved? How do we know that we have reached the limit that a particular technique offers? To these questions we can also add questions concerning the selection of the best descriptors from a large pool of currently available molecular descriptors; questions concerning the structural interpretation of combinations of descriptors (given by the regression equation); and questions concerning alternative choices of descriptors that show a similar statistical characteristic, i.e., similar correlation coefficient (r), similar standard error (s), and similar Fisher ratio (F).

In this paper we will investigate structure−property relationships by close examination of BP of alcohols and by using a simple novel molecular descriptor which apparently has been hitherto overlooked. The new descriptor may be of interest also in other QSPR and QSAR studies. Before we outline the approach we will try to answer why a regression equation with satisfactory statistical parameters already reported should be reconsidered.

ANSWERS

There are several reasons why an improvement of a regression analysis is important and desirable:

(1) In the rigorous sciences the highest level of accuracy and precision are important, whether it concerns the third or fifth decimal place, whether it applies to experimental measurements or theoretically computed quantities. This has been recognized by most quantum chemists. If that would not be so, we would still today describe an H_2 molecule the way Heitler and London outlined in 1927[5] and not the way how Kolos and Wolniewicz did about 40 years later.[6]

(2) Improved accuracy of a regression may point to "hidden" experimental errors in the raw data or "unusual" behavior of a compound that distinguishes itself from the rest. Such a distinction may be due to unrecognized structural features that are beyond the ability of the molecular descriptors to capture. An illustration of a "hidden" experimental error has recently been pointed out by correlating two experimental solubilities of alcohols (log S with ln γ).[7] An outlier (belonging to 2-hexanol) was visible only because the two properties showed exceedingly high correlation. Unfortunately we do not know which of the two solubilities is in error. If we would have molecular descriptors that could give correlation with properties of a similar high quality, then we could determine whether log S or ln γ is in error.

(3) Sometimes even a small improvement of a regression may call for reinterpretation of the results of the analysis. For example, use of the simple connectivity index[8] in a correlation of the toxicity of aliphatic ethers in mice[9] gave a fair regression, characterized by $r = 0.9548$, $s = 0.130$, and $F = 93$. For this case bond contributions to the connectivity index of individual CO bonds and CC bond are equal. If, however, a variable connectivity index[10,11] is employed, then the optimal regression gives somewhat better statistics: $r = 0.9756$, $s = 0.096$, and $F = 178$. However, the optimal descriptors show that the contributions of CO bonds are now about 30 times larger than those of similar CC bonds. Thus, although there was no dramatic change in the statistical parameters of the regression, the interpretation of the descriptors has dramatically changed. This kind of

information would be lost if one would be satisfied with inferior regressions, even if of acceptable quality.

(4) Search for a better regression may facilitate interpretation of available results. An alternative regression may (a) point to molecular descriptors that have simpler structural interpretation; (b) express a correlation with fewer descriptors (which may make interpretation simpler); (c) express a correlation by a simpler mathematical form (e.g., by a linear instead of a polynomial regression); (d) express a correlation with significantly better statistics (e.g., regression accompanied by visibly reduced standard error); and (e) express a correlation by novel structural descriptors hitherto overlooked.

To illustrate this last point consider the quadratic regression for octane motor numbers using the leading eigenvalue of the adjacency matrix λ_1 as descriptor.[12] It gives the following statistical parameters: $r = 0.9648$, $s = 6.98$, and $F = 87$. The leading eigenvalue was interpreted by Lovasz and Pelikan as a branching index.[13] Satisfactory regression is to be expected since more branched octanes have larger octane numbers. A result of similar quality has also been obtained using molecular ID numbers and the Wiener number as descriptors in quadratic correlations, which gave $r = 0.9689$, $s = 6.56$, and $F = 100$ and $r = 0.9711$, $s = 6.33$, and $F = 108$, respectively. Although ID and W produce slightly better regressions, interpretation of ID and W is less straightforward.

Could the above results be visibly improved? Not long ago the path matrices[14−16] were introduced as a source of novel structural invariants. When in constructed path matrices the leading eigenvalue of path replace paths as matrix elements one obtains a numerical matrix. The leading eigenvalue of so constructed matrices, $\lambda\lambda_1$, represents a novel molecular descriptor. The $\lambda\lambda_1$ is related to the leading eigenvalue of the adjacency matrix λ_1. The regression between λ_1 and $\lambda\lambda_1$ has for the coefficient of regression 0.9820 ($s = 0.0257$ and $F = 378$). Hence, both descriptors may be viewed as alternative branching indices. When $\lambda\lambda_1$ is used in a quadratic regression, we obtain the following statistics: $r = 0.9891$, $s = 3.91$, and $F = 292$.

The numerical differences between λ_1 and $\lambda\lambda_1$ for individual octane isomers are small indeed, yet the standard error associated with $\lambda\lambda_1$ has been dramatically reduced, almost by half. Thus we see that even minor changes among two descriptors may be very important, if the differences reflects the relevant parts of the second descriptor.

Can this very good result be further improved? Are there alternative one-variable regressions of similar quality? Why do the leading eigenvalues of the adjacency matrix and the path matrix reflect molecular branching? We will see in the next section that indeed we can improve upon the best current regression for motor octane numbers. Use of alternative descriptors that give the same statistical parameters has been discussed elsewhere.[17] Such alternative regressions are of considerable importance when one is interested in interpretation of the results obtained.

OPTIMAL MOLECULAR DESCRIPTORS

Once a molecule is selected, the numerical value for almost all molecular descriptors, including several hundred topological indices can be calculated. We may refer to such descriptors as "fixed" in contrast with a few recent descriptors that involve a variable part, which has yet to be determined when a particular application is considered. A "fixed" molecular descriptor, once calculated, holds for any applications in which the particular molecule is involved. In contrast, the numerical value of "variable" molecular descriptors may vary for the same molecule from application to application. In general these "variable" descriptors may be sensitive not only to the number of other descriptors used but also to the molecular property considered.

Typically "variable" descriptors associate an unspecified weight with a heteroatom or bond. This weight is systematically varied till the minimal standard error for the particular structure−property correlation is obtained. Such a "variable" descriptor was introduced over 10 years.[10,11] By varying the variable x associated with the oxygen atom of alcohols it was possible to reduce the standard error for the boiling points of hexanols and heptanols (when a single molecular descriptor was used) by more than half, from the value 7.86 °C to 3.83 °C.

The improvement of a correlation in which the standard error can be decreased by half can be characterized as dramatic. It clearly shows that the differentiation between carbon and oxygen atoms in alcohols is essential for the correlation of the particular property. The above result has been further improved by introducing variable weights for both the carbon atom and the oxygen atom. It was found that the values $x_C = 1.50$ and $x_O = -0.85$ give the minimal standard error for the boiling points, $s = 3.30$. As has been further outlined in ref 11 if one uses multiple regression that employs variable index $^2\chi$, besides the (variable) connectivity index $^1\chi$, one can further reduce the standard error to $s = 2.29$ °C, which is a reduction by a full one degree Celsius.

Use of two descriptors, each of which has two variable components (optimized for carbon and oxygen atoms), has not only resulted in very respectable correlations with the standard error slightly above 2 °C but also made it possible to identify potential outliers. After eliminating 2,3-dimethyl-1-butanol as outlier for the remaining 16 hexanols one finds the correlation coefficient $r = 0.992$ and the standard error well below 2 °C.

Can this result be improved? Have we reached the limit that the model (multiple regression based on the weighted connectivity index and closely related weighted paths of length two) can support? Are we missing a descriptor that can yield even better results?

HOW TO IMPROVE A REGRESSION

An exhaustive search for a better combination of molecular descriptors is tedious and not necessarily practical when one has to screen a very large pool of descriptors. It would be better if there was a more "directed" way to find or even search for "missing" descriptors. But how can we "design" a more "directed" search for a missing descriptor? One possible route is to even further generalize already generalized descriptors. We will illustrate this by considering the boiling points of alcohols using, instead of the weighted connectivity indices, weighted paths.[7,18−20]

WEIGHTED PATHS

Randić and Basak have recently reexamined the correlation of the boiling points for a set of 58 alcohols (from methanol

Table 3. Contributions to the Variable Connectivity Index According to Vertex Degrees of Bond Endpoints (Upper Part) and to Path Numbers by Different Bonds in Smaller Alcohols (Lower Part)

	1-butanol	2-butanol	2-Me-1-propanol	2-Me-2-propanol
p_1	$3+x$	$3+x$	$3+x$	$3+x$
p_2	$2+x$	$2+2x$	$3+x$	$3+3x$
p_3	$1+x$	$1+x$	$2x$	0

Table 4. Comparison of Several Reported Regressions for the Boiling Points of Alcohols

n	descriptors	*r*	*s*	ref
	One-Variable			
37	$^1\chi^v$	0.9555	10.3	*a*
62[g]	log ξ	0.95	9.25	*b*
62[g]	log W	0.92	11.89	*b*
17 hexanols	$^1\chi$		7.86	*c*
17	variable $^1\chi$		3.30	*c*
37 heptanols	variable $^1\chi$	0.944	3.91	*c*
37	$^1\chi$, $^2\pi$ variable	0.963	3.25	*c*
62	one descriptor	0.9459	9.66	*d*
	Two-Variables			
37	$^1\chi$, $^1\chi^v$	0.9924	4.4	*a*
17	$^1\chi$, $^2\pi$ variable	0.985	2.29	*c*
16[h]	$^1\chi$, $^2\pi$ variable	0.992	1.70	*c*
62	two descriptors	0.9943	3.12	*d*
123	two descriptors	0.975	4.05	*e*
	Three-Variables			
37	N_C, T_m, C_α	0.9962	3.13	*f*
37	W_r, P, s_{ox}	0.9915	4.70	*a*
62	three descriptors	0.9976	2.05	*d*
123	three descriptors	0.982	3.49	*e*

[a] Reference 29. [b] Sharma, V.; Goswami, R.; Madan, A. K. Eccentric connectivity index: A novel highly discriminating topological descriptor for structure−property and structure−activity studies. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 273−282. [c] Randić, M. On computation of optimal parameters for multivariate analysis of structure−property relationship. *J. Comput. Chem.* **1991**, *12*, 970−980. [d] Kier, L. B.; Hall, L. H. Molecular connectivity VII. Specific treatment of heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809. [e] Smeeks, F. C.; Jurs, P. C. Prediction of boiling points of alcohols from molecular structure. *Anal. Chim. Acta* **1990**, *233*, 111−119. [f] Recalculated using the data from ref 29. When compounds #16 and # 30 are discarded as outliers the correlation coefficient becomes 0.9978 and the standard error $s = 2.46$ °C. This is still much higher than the value quoted in ref 29 for $s = 1.44$ °C. [g] Correlation with log BP. [h] Outlier 2,3-dimethyl-1-butanol is discarded.

to 1-decanol) using weighted paths as molecular descriptors.[19] They introduced variable weights for paths involving oxygen atoms. In Table 3 we illustrate the weighted paths of length one, two, and three for 1-butanol, 2-butanol, 2-methyl-1-propanol, and 2-methyl-2-propanol by assigning the weight x to the CO bond only. In Table 4 we have summarized previous results on the correlation of the boiling points of alcohols from different investigators in order to illustrate the standard errors obtained. We should emphasize, however, that different authors used somewhat different sets of compounds and different numbers of descriptors. Hence, the comparison of results in Table 4 does not necessarily reflects subtleties of individual studies.

Computationally the technique closest to what we will elaborate here was the work based on weighted paths used as molecular descriptors.[7,18−20] A simple regression using only paths of length one gives for the standard regression $s = 13.28$ °C. Because the weight x is strictly additive for all molecules its variation cannot change statistical parameters in a linear regression. Thus the above is the worst possible result that a single path descriptor can give. What is the best result that paths can give? Using two descriptors (p_1 and p_2) the standard error was reduced to 6.64 °C when no weight was used. If now the weight x is varied (and this only affects p_2, not p_1 as already pointed out), it was found that when $x = 2.6$ the standard error dropped to 4.04 °C. The accompanying regression coefficient is $r = 0.9938$ and the Fisher ratio, a measure of the quality of regression, is $F = 2193$. By using three descriptors, p_1, p_2, and p_3, one finds that the accompanying regression coefficient becomes slightly better, $r = 0.9943$, the standard error has improved slightly, $s = 3.89$ °C, and the Fisher ratio has somewhat decreased, $F = 1578$, while the optimal value of x has slightly shifted ($x = 3.1$). If we compare the above standard error with those listed in Table 4, we see that this result is among the best. Can this result be further improved? In view of the fact that we may be near the limit of improvements, any betterment (if possible) is not likely to be dramatic. Hence, a further reduction in the standard error of half a degree of Celsius may represent a challenge.

MISSING DESCRIPTOR

Let us try to further generalize already generalized descriptors. When path numbers are used as descriptors there is no need to consider different weights for carbon atoms and oxygen atoms because one can always factor out one of the variable weights (say y). This will reduce the other weight to a novel weight x/y (for oxygen) leaving carbons without variable weight. Such a procedure does not affect linear regression, just as an addition of x to p_1 did not change the role of p_1 as descriptor. What else we can do?

We can vary separately the weights for p_2 and p_3 and try to find optimal weights x (for p_2) and y (for p_3) that will reduce the standard error. This seems a plausible idea because we have already noticed that when a single weight x is used first for p_2 separately and then both for p_2 and p_3, we obtained somewhat different values of x for the two cases. The same has been observed for the "variable" connectivity indices.[21] In Table 5 we give the standard statistical parameters (r, s, and F) for correlation of the boiling points of alcohols using p_1, p_2, and p_3 for various values of y (for p_3) using the fixed the optimal value for x ($x = 3.1$, for p_2). As can be seen at the top of Table 5, as we increased the value of y the standard error went down (and other two statistical parameters improved). Table 5 shows the actual search for best values of y by keeping a fixed value for $x = 3.1$ (the weight of p_2). By increasing y we see a gradual but slow decrease in the standard error. By continuing the search we had to consider larger and larger values of y in the hope that the minimum will be reached, i.e., the next successive value will exceed the preceding. However, no minimum was found, and as y increased the standard error continued to decrease at each successive step. By reaching $y = 100$ one starts to suspect that the decrease of the standard error could continue indefinitely—as indeed it does. The statistical parameters corresponding to the limiting value $y =$ infinity are shown in the last column of Table 5. Indeed, we see that the corresponding standard error is the smallest, the correlation coefficient is the largest, and also that the Fisher ratio is the largest. Thus the value $y = \infty$ is the optimal choice for the weight for paths of length three.

Table 5. Variation of the Statistical Parameters *r*, *s*, and *F* (the Coefficient of the Regressions, the Standard Error, and the Fisher Ratio) for the Boiling Points of 58 Alcohols Studied When Variable *y* Increases

y	2.7	3.5	4	5	6	8	10	15	20	30	40	50	100	1000	10000	infinity
r	0.99426	0.99441	0.99449	0.99459	0.99466	0.99476	0.99481	0.99486	0.99493	0.99496	0.99498	0.99499	0.99502	0.99504	0.99504	0.99504
s	3.920	3.867	3.842	3.805	3.780	3.744	3.727	3.700	3.686	3.674	3.665	3.661	3.653	3.645	3.64438	3.64430
F	1554	1597	1619	1650	1673	1705	1721	1745	1760	1771	1780	1784	1793	1800.87	1800.87	1800.96

Table 6. Variation of the Statistical Parameters *r*, *s*, and *F* (the Coefficient of the Regressions, the Standard Error, and the Fisher Ratio) for the Boiling Points of 58 Alcohols Studied When Variable *x* Varies and *y* Is Assumed at the Limit

$x = 2$	$x = 2.5$	$x = 2.7$	$x = 2.8$	$x = 3$	$x = 3.1$	3.2	3.5
0.99362	0.99473	0.99492	0.99497	0.99503	0.99504	0.99504	0.99497
4.132	3.755	3.689	3.669	3.647	3.644	3.646	3.668
1397	1695	1757	1778	1798	1801	1800	1777

It was observed when testing the initial values of *x* and *y* that a small change in *x* causes but a minor change of the statistical parameters of the regression. For example, when $x = 3.0$ (instead of $x = 3.1$ that is shown in Table 5) we have $r = 0.99442$, $s = 3.864$, and $F = 1600$ instead of $r = 0.99441$, $s = 3.867$, and $F = 1597$. Thus the sensitivity of the analysis is dominated by *y*, rather than *x*. Nevertheless we have to find the best value of *x* when $y = \infty$. In Table 6 we illustrate the search when *x* is varied as *y* is kept constant (at infinity). As we can see, the smallest standard error is found when $x = 3.1$, confirming thus that indeed this is the optimal weight for p_2.

PATHS OF INFINITE WEIGHT

What is the significance of *y* becoming infinity? Clearly, that simply means that when one considers paths of length three the only contributions that are relevant are the paths of length three originating from the oxygen atom. Therefore, instead of considering weighted paths of length two and three and searching for optimal weights for *x* and *y* (what we just did) we could have (if we knew this in advance) considered a regression based on the weighted paths of length two and as an additional descriptor considered the count of paths of length three originating from the oxygen atom. The result would be precisely the same. Hence, the number of paths of length three originating from the oxygen atom represents a novel topological index, p_3*, the "missing" descriptor that makes the best regression for the boiling points of alcohols when variable paths are used as descriptors. If we compare such a regression with the best regression using weighted p_1, p_2, and p_3, we see that the standard error of 3.89 °C now has been reduced to 3.64 °C. This at first may not appear much, but recollect that the value of 3.89 °C is already quite a good result. In addition, look at the simplicity of our descriptors: both p_1 and p_3* are represented by integers, only p_2 is a variable descriptor. Moreover, we can interpret the results in simpler structural concepts than most of those summarized in Table 4.

Platt already in 1947 suggested paths of different length as molecular descriptors for discussion of variation of molecular properties among isomers.[22] Paths of length three have been one of the two descriptors that already Wiener recognized in 1947 as critical for structure−property correlation.[23] He called it "polarity" index, P, but this index is related to "steric hindrance" of atoms or "compactness" as Wiener also referred to it. One can easily verify that p_3 increases for more branched paraffins by comparing the path numbers for heptane, octane, and nonane isomers (listed in ref 25). It is interesting here also to recall that paths of length three, but only those between terminal carbon atoms, have been recognized as contributing significantly to improving regression of chromatographic retention indices.[24] Here, we have an important new descriptor, which we will label as p_3*. Again it is based on the count of paths of length three; however, paths associated only with the oxygen atom in monohydroxylic alcohols. Clearly p_3* can therefore be interpreted as characterizing "steric hindrance" or "compactness" around the oxygen atom in alcohols, which is a novel critical structural element that has been hitherto overlooked. Kier and Hall[25] noticed that the molecular connectivity index $^3\chi$ parallels the number of *guache*−*trans* rearrangements in a molecule and carries similar information to p_3. On the basis of these considerations they arrived at their flexibility index. We should also add that for aliphatic alcohols Seybold, May, and Bagel[25] employed a convenient additional index C_α which is the number of carbon atoms bonded to the alpha carbon atom. This index, according to Seybold et al., accounts for steric hindrance and improves correlations for the boiling points in alcohols and few other properties. When combined with descriptors N_C, the number of carbon atoms, which can be taken as a measure of size, and T_m, the number of terminal methyl groups, which can be taken as a measure of branching or compactness, it gives an excellent three-variable regression with $r = 0.996$ and $s = 1.44$. The index C_α would be in our notation p_2*, while we used as an index that measured steric hindrance around oxygen index p_3*.

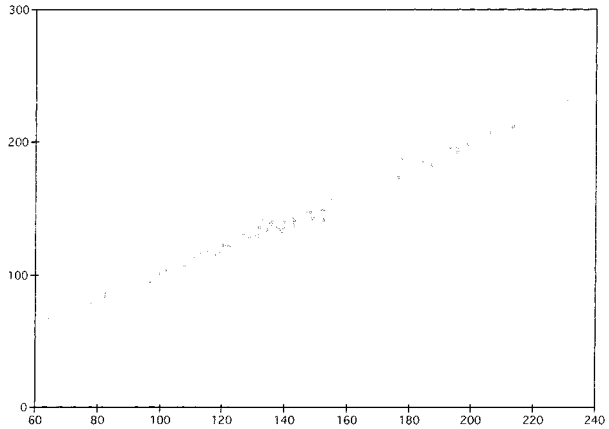

Figure 1. The computed against the experimental boiling points for 58 alcohols.

OUTLIERS

Did we exhaust all reasonable possibilities for improving of the regression for the boiling points for the set of 58 alcohols and reach the "end" of what MRA can offer? In Figure 1 we show the plot of the calculated BPs against the experimental ones. The accompanying stepwise regression equations (with the probability estimates for the coefficients

Table 7. Stepwise Regression Equations for the Boiling Points of 58 Alcohols Using Weighted Paths p_1, p_2 and the Limiting Value p_3*[a]

p_1	p_2	p_3	const	r	s	F
			BP			
17.6576			**−5.2642**	0.9294	13.28	355
23.7595	**−4.6212**		−6.6367	0.9935	4.13	2101
23.7305	−4.2632	**−2.3294**	−6.2533	0.9950	3.64	1801
17.6576	**−4.6212**	**−2.3294**	**−5.2642**	**0.9950**	**3.64**	**1801**
p_1	p_2	p_3*	const	r	s	F
			BP*			
18.1269			**−8.1578**	0.9370	12.79	381
23.8988	**−4.6544**		−7.2339	0.9953	3.59	2738
23.9188	−4. 3235	**−2.5158**	−6.7248	0.9970	2.89	2826
18.1269	**−4.6544**	**−2.5158**	**−8.1578**	**0.9970**	**2.89**	**2826**

[a] The "diagonal" entries (the coefficient of the last introduced descriptor at each step) are shown in boldface. These entries form the regression equation corresponding to orthogonalized descriptors (shown as a separate line in the part of the table giving BP and BP*, in which three outliers have been removed).

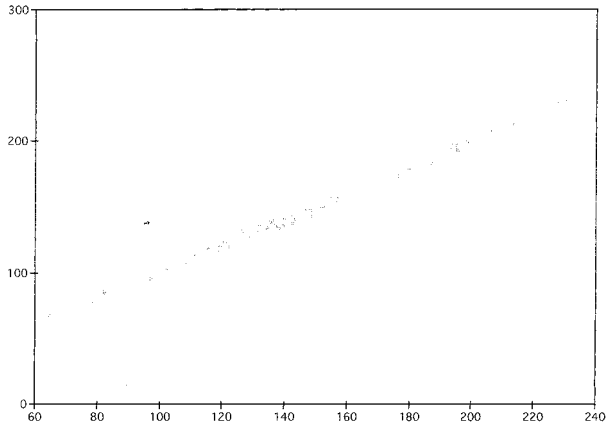

Figure 2. The computed against the experimental boiling points for 55 alcholols (the outliers removed and their residuals are as follows: 2,2,3-trimethyl-3-pentanol (+10.37 °C), 2,6-dimethyl-4-heptanol (−9.70 °C), and 3,3-dimethyl-2-pentanol (−9.17 °C).

of the regression equation) are listed in Table 7, while the computed BP and the residuals are shown in Table 8. The "worst" compounds and their residuals are listed as follows:

2,2,3-trimethyl-3-pentanol +10.37 °C

2,6-dimethyl-4-heptanol −9.70

3,3-dimethyl-2-pentanol −9.17

With the standard error of 3.64 °C clearly these three compounds can be identified as outliers. When we remove these three compounds from the list we obtain the regression equations shown in the lower part of Table 7. Observe the dramatic reduction of the standard error from the value of 3.64 °C to well below 3 °C (s = 2.89 °C). The plotted computed BP (now designated as BP*) against the experimental BP is shown in Figure 2 where inspection can discern even visually an improvement of the regression. The actual computed BP* and the accompanying residuals (res*) are shown in the last two columns of Table 8. Observe also the dramatic increase of the Fisher ratio (which almost doubled) upon exclusion of the three outliers, which supports elimination of the three as outliers. In Table 7 in bold format are indicated the coefficients of the stepwise regression which appear in the orthogonalized regression,[26,27] which has also been shown as the last row corresponding to BP and BP* parts of the table.

Table 8. Experimental and Computed Boiling Points and the Residuals for the 58 Alcohols Listed in Ref 20[a]

	compound	BP_{exp}	BP_{calc}	res	BP^*_{calc}	res*
1	methanol	64.7	67.31	−2.61	67.42	−2.72
2	ethanol	78.3	77.83	0.47	77.94	0.36
3	1-propanol	97.2	94.96	2.24	95.02	2.18
4	2-propanol	82.3	84.08	−1.78	84.13	−1.83
5	1-butanol	117.7	114.43	3.27	114.61	3.09
6	2-butanol	99.6	101.22	−1.62	101.21	−1.61
7	2M-1-propanol	107.9	107.84	0.06	107.77	0.13
8	2M-2-propanol	82.4	86.07	−3.67	86.00	−3.60
9	1-pentanol	137.8	133.90	3.90	134.21	3.59
10	2-pentanol	119.0	120.68	−1.68	120.81	−1.81
11	3-pentanol	115.3	118.35	−3.05	118.29	−2.99
12	2M-1-butanol	128.7	127.31	1.39	127.37	1.33
13	3M-1-butanol	131.2	129.64	1.56	129.89	1.31
14	2M-2-butanol	102.0	103.20	−1.20	103.08	−1.08
15	3M-2-butanol	111.5	114.09	−2.59	113.97	−2.47
16	2,2MM-1-propanol	113.1	116.45	−3.35	116.21	−3.11
17	1-hexanol	157.0	153.37	3.63	153.80	3.20
18	2-hexanol	139.9	140.15	−0.25	140.40	−0.50
19	3-hexanol	135.4	137.82	−2.42	137.89	−2.49
20	2M-1-pentanol	148.0	146.77	1.23	146.97	1.03
21	3M-1-pentanol	152.4	149.10	3.30	149.48	2.92
22	4-M-1-pentanol	151.8	149.10	2.70	149.48	2.32
23	2M-2-pentanol	121.4	122.67	−1.27	122.68	−1.28
24	3M-2-pentanol	134.2	133.56	0.64	133.56	0.64
25	4M-2-pentanol	131.7	135.89	−4.19	136.08	−4.38
26	2M-3-pentanol	126.5	131.23	−4.73	131.05	−4.55
27	3M-3-pentanol	122.4	120.34	2.06	120.16	2.24
28	2E−1-butanol	146.5	146.77	−0.27	146.97	−0.47
29	2,2MM-1-butanol	136.8	135.92	0.88	135.80	1.00
30	2,3MM-1-butanol	149.0	142.51	6.49	142.64	6.36
31	3,3MM-1-butanol	143.0	140.58	2.42	140.83	2.17
32	2,3MM-2-butanol	118.6	116.08	2.52	115.84	2.76
33	3,3MM-2-butanol	120.0	122.70	−2.70	122.40	−2.40
34	1-heptanol	176.3	172.83	3.47	173.40	2.90
35	3-heptanol	156.8	157.29	−0.49	157.48	−0.68
36	4-heptanol	155.0	157.29	−2.29	157.48	−2.48
37	2M-2-hexanol	142.5	142.14	0.36	142.27	0.23
38	3M-3-hexanol	142.4	139.81	2.59	139.76	2.64
39	3E−3-pentanol	142.5	137.48	5.02	137.24	5.26
40	2,3MM-2-pentanol	139.7	135.55	4.15	135.43	4.27
41	3,3MM-2-pentanol	133.0	142.17	−9.17		
42	2,2MM-3-pentanol	136.0	139.84	−3.84	139.48	−3.48
43	2,3MM-3-pentanol	139.0	133.22	5.78	132.92	6.08
44	2,4MM-3-pentanol	138.8	144.10	−5.30	143.80	−5.00
45	1-octanol	195.2	192.30	2.90	193.00	2.20
46	2-octanol	179.8	179.08	0.72	179.59	0.21
47	2E−1-hexanol	184.6	185.71	−1.11	186.16	−1.56
48	2,2,3MMM-3-pentanol	152.2	141.83	10.37		
49	1-nonanol	213.1	211.77	1.33	212.59	0.51
50	2-nonanol	198.5	198.55	−0.05	199.19	−0.69
51	3-nonanol	194.7	196.22	−1.52	196.67	−1.97
52	4-nonanol	193.0	196.22	−3.22	196.67	−3.67
53	5-nonanol	195.1	196.22	−1.12	196.67	−1.57
54	7-M-1-octanol	206.0	207.50	−1.50	208.27	−2.27
55	2,6MM-3-heptanol	178.0	187.70	−9.70		
56	3,5MM-4-heptanol	187.0	183.04	3.96	182.99	4.01
57	3,5,5MMM-1-hexanol	193.0	194.71	−1.71	195.30	−2.30
58	1-decanol	230.2	231.24	−1.04	232.19	−1.99

[a] The last two columns correspond to data from which three outliers have been removed.

CONCLUDING REMARKS

We have raised several questions regarding multivariate regression analysis and have illustrated (using correlation of the boiling points of alcohols and employing weighted path

numbers p_1, p_2, and p_3) how some of the questions posed can be answered. In particular, we have seen how generalization of existing topological indices by considering variable weights can lead to novel descriptors. In this way we arrived at novel descriptor p_3*, the count of paths of length three for the heteroatom only. The novel descriptor has a clear structural interpretation: it characterizes the "steric hindrance" around the heteroatom (here oxygen), which apparently is an important structural feature of alcohols. The new descriptor plays a similar role to C_α descriptor introduced by Seybold, May, and Bagel,[29] with the distinction that we have arrived at p_3* by a constructive search, rather than selecting a descriptor in an ad hoc manner, as was the case with C_α. However, it is interesting to see that the three simple indicator-descriptors, N_C, T_m, and C_α, nevertheless appear to be the best descriptors in the case of the boiling points of alcohols.

ACKNOWLEDGMENT

This is contribution number 288 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by Grants F49620-98-1-0015 and F49620-01-1-0098 from the United States Airforce. We thank Professor A. T. Balaban for critical review of this manuscript and his many comments that improved the manuscript.

REFERENCES AND NOTES

(1) Kohn, M. C. Strategies for computer modeling. *Bull. Math. Biol*. **1986**, *48*, 417−426.
(2) Basak, S. C.; Niemi, G. J.; Veith, G. D. Predicting properties of molecules using graph invariants. *J. Math. Chem*. **1991**, *7*, 243−272.
(3) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992.
(4) Randić, M.; Zupan, J. On the interpretation of well-known topological indices. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 550−560.
(5) Heitler, W.; London, F. Wechselwirkung neutraler Atome und homopolare Bindung nach der Quantenmechanik. *Zeit. Phys*. **1927**, *44*, 455−472.
(6) Kolos, W.; Wolniewicz, L. J. Accurate adiabatic treatment of the ground state of the hydrogen molecule. *J. Chem. Phys*. **1964**, *41*, 3663−3673.
(7) Randić, M.; Basak, S. C. Multiple regression analysis with optimal descriptors. *SAR QSAR Environ. Res*. **2000**, *11*, 1−23.
(8) Randić, M. On the characterization of molecular branching. *J. Am. Chem. Soc*. **1975**, *97*, 6609−6615.
(9) Randić, M.; Basak, S. C. On use of the variable connectivity index $^1\chi^f$ in QSAR: Toxicity of aliphatic ethers. *J. Comput. Chem*. **2001**, *41*, 614−618.
(10) Randić, M. Novel graph theoretical approach to heteroatom in quantitative structure−activity relationship. *Chemometrics Intel. Lab. Syst*. **1991**, *12*, 970−980.
(11) Randić, M. On computation of optimal parameters for multivariate analysis of structure−property relationship. *J. Comput. Chem*. **1991**, *12*, 70−980.
(12) Randić, M. On molecular branching. *Acta Chim. Slov*. **1997**, *44*, 57−77.
(13) Lovasz, L.; Pelikan, J. On the eigenvalues of trees. *Period. Math. Hung*. **1973**, *3*, 175−182.
(14) Randić, M.; Plavsic, D.; Razinger, M. Double invariants. *MATCH* **1997**, *35*, 243−259.
(15) Randić, M. On structural ordering and branching of acyclic saturated hydrocarbons. *J. Math. Chem. 24*, 345−358.
(16) Randić, M.; Guo, X.; Bobst, S. Use of path matrices for a characterization of molecular structures. *DIMACS Ser. Discrete Mathematics Theoretical Comput. Sci*. **2000**, *51*, 305−322.
(17) Randić, M. Linear combinations of path numbers as molecular descriptors. *New J. Chem*. **1997**, *21*, 945−951.
(18) Randić, M.; Pompe, M. On characterization of CC double bond in alkenes. *SAR QSAR Environ. Res*. **1999**, *10*, 451−471.
(19) Randić, M.; Basak, S. C. Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 261−266.
(20) Randić, M.; Pompe, M. The variable molecular descriptors based on distance related matrices. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 575−581.
(21) Randić, M.; Basak, S. C. Variable molecular descriptors. In *Some Aspects of Mathematical Chemistry*; Sinha, D. K., Basak. S. C., Mohanty, R. K., Busamallick, I. N., Eds.; Visva-Bharati University Press: Santiniketan, India, In press.
(22) Platt, J. R. Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys*. **1947**, *15*, 419.
(23) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc*. **1947**, *69*, 17−20.
(24) Randić, M.; Wilkins, C. L. Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data. *J. Chem. Phys*. **1979**, *83*, 1525−1540.
(25) Randić, M. On structural origin of chromatographic retention data. *J. Chromatogr*. **1978**, *161*, 1−14.
(26) Kier, L. B.; Hall, L. H. Structural information and flexibility index from the molecular connectivity $^3\chi_p$ index. *Quant. Struct.-Act. Relat*. **1983**, *2*, 55−59.
(27) Randić, M. Orthogonal molecular descriptors. *New J. Chem. 15*, 517−525.
(28) Randić, M. Resolution of ambiguities in structure−property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 311−320.
(29) Seybold, P. G.; May, M.; Bagal, U. A. Molecular structure − Property relationship. *J. Chem. Educ*. **1987**, *64*, 575−581.

CI000116E