Graph Valence Shells as Molecular Descriptors

Milan Randić[†]

National Institute of Chemistry, Ljubljana, Slovenia, and Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311

Received August 12, 2000

We have introduced a new simple structural descriptor for molecules that is based on the count of the valence shells for vertices in molecular graphs. The construction of the new descriptor is illustrated on 2,3-dimethylhexane and is reported for the 18 octane isomers. The relationship of the new descriptor to the path numbers of a graph is discussed. It can be seen that the path counts and the count of valence of neighbor shells are related for paths of length two (and shells of range two). There is no appreciable correlation between the count of the longer paths and the count of the corresponding neighbor valence shells at larger separations. Use of the neighbor valence shells as molecular descriptors is illustrated on the boiling point, the entropy, and the density of octanes. An intriguing situation is observed for regressions involving considered properties of *n*-octane isomers C_8H_{18} in that the paths of length two, three, and four and the shells of the range two, three, and four give *identical* multivariate regression statistics. An explanation for this somewhat unusual aspect of MRA (multiple regression analysis) is offered.

INTRODUCTION

Multiple regression analysis (MRA), one of the oldest data reduction methodologies, continues to be widely used in QSPR and QSAR (the quantitative structure-property relationship and the quantitative structure-activity relationship, respectively). As has been recognized for some time, a success or a failure of a structure-property-activity study often critically depends on the selection of molecular descriptors. Hence, it is not surprising to see continual development of novel molecular descriptors.[1] In this contribution we will introduce a set of novel molecular descriptors with a simple structural interpretation, which unexpectedly shows an intriguing relationship to path numbers, p_k.

PATH NUMBERS

Paths and walks represent the most elementary graph theoretical (structural) concepts. The path of length k, p_k, is defined as a sequence of k consecutive edges of a structure such that no vertex and no edge is repeated in the sequence. The walk of length k, w_k, is defined as a sequence of k incident edges in which both the vertices and the edges could be repeated any number of times. When no confusion is likely to arise, we will use symbols p_k and w_k to signify both the paths and walks of length k, respectively, as well as the number of paths and the number of walks of length k. In Table 1 (the left side) we illustrate the count of paths of increasing length for 2,3-dimethylhexane.

The path numbers were suggested as potentially useful molecular descriptors already 50 years ago by Platt.[2] However, it was only with the revival of the chemical graph theory in 1970s that characterization of molecules by paths received due attention.[3] Because one can view paths and walks as the most elementary structural invariants recently Randić and Zupan[4] suggested the use of paths and walks for interpretation of more involved topological indices. In particular they have shown how the Hosoya Z topological index,[5] the Wiener number[6] W, and the connectivity index[7] $^1\chi$ can be decomposed in bond contributions (i.e., paths of length one).

There are hundreds of topological indices that have been introduced in the past two decades and applied in QSPR and QSAR studies. For example, the computer program CODESSA[8] evaluates about 400 molecular descriptors (topological, geometrical and quantum chemical), but, with the exception of a few, a majority of molecular descriptors are defined by relatively involved mathematical expressions or contain empirical parameters. For a recent compilation of strategies to develop novel graph-theoretic descriptors one should consult a chapter by Estrada in a recent book on the use of topological indices in QSAR and QSPR.[9] For interpretation of the MRA results direct use of structurally simple descriptors seems desirable, or alternatively if mathematically more involved descriptors are used simple descriptors can be used indirectly to interpret more involved descriptors. Hence, interpretation of MRA would be facilitated if we could increase the number of simple structural descriptors that can serve as elementary structural invariants. In this contribution we propose one such set of descriptors which may be of interest for structure−property-activity studies.

VALENCE SHELL COUNT

The concept of neighbor shells is similar to the concept of paths. The difference is that instead of counting for each atom the number of neighbors at increasing length, one adds the valences of neighbors at increasing separation. In Table 1 (the right side) we illustrate the count of valence shells for carbon atoms of 2,3-dimethylhexane (shown in Figure 1). A shell of range zero represents the valence of a vertex,

[†] Corresponding author phone: (515)292-7411; fax: (515)292-8629; e-mail: milan.randic@drake.edu. Current address: 3225 Kingman Rd., Ames, IA 50014.

10.1021/ci000121i CCC: $20.00
Published on Web 04/05/2001

Table 1. The Count of Paths and the Count of Shells of Increasing Length or Range in 2,3-Dimethylhexane

atom	p_1	p_2	p_3	p_4	p_5	atom	s_0	s_1	s_2	s_3	s_4	s_5
1	1	2	2	1	1	1	1	3	4	3	2	2
2	3	2	1	1	0	2	3	5	3	2	1	1
3	3	3	1	0	0	3	3	6	4	1		
4	2	3	2	0	0	4	2	5	5	2		
5	2	1	2	2	0	5	2	3	3	4	2	2
6	1	1	1	2	2	6	1	2	2	3	4	4
7	1	2	2	1	1	7	1	3	4	3	2	2
8	1	2	3	1	0	8	1	3	5	4	1	1
molecule: 7, 8, 7, 4, 2						molecule: 7, 15, 15, 11, 6, 2						

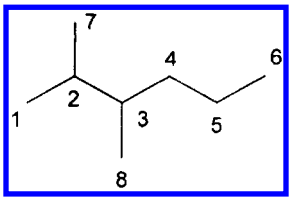

Figure 1. Numbering of carbon atoms in a molecular graphs of 2,3-dimethylhexane.

Table 2. The Paths Counts p_2, p_3, p_4 and the Valence Shell Counts s_2, s_3, s_4 for the 18 Isomers of Octane C_8H_{18}

octane isomer	p_2	p_3	p_4	s_1	s_2	s_3
n-octane	6	5	4	13	11	9
2-methylheptane	7	5	4	14	12	9
3-methylheptane	7	6	4	14	13	10
4-methylheptane	7	6	5	14	13	11
3-ethylhexane	7	7	5	14	14	12
2,2-dimethylhexane	9	5	4	16	14	9
2,3-dimethylhexane	8	7	4	15	15	11
2,4-dimethylhexane	8	6	5	15	14	11
2,5-dimethylhexane	8	5	4	15	13	9
3,3-dimethylhexane	9	7	4	16	16	11
3,4-dimethylhexane	8	8	4	15	16	12
2-methyl-3-ethylpentane	8	8	5	15	16	13
3-methyl-3-ethylpentane	9	9	3	16	18	12
2,2,3-trimethylpentane	10	8	3	17	18	11
2,2,4-trimethylpentane	10	5	6	17	15	11
2,3,3-trimethylpentane	10	9	2	17	19	11
2,3,4-trimethylpentane	9	8	4	16	17	12
2,2,3,3-tetramethylbutane	12	9	0	19	21	9

while a shell of range one gives the so-called extended valence. One may view the count of shells as the count of weighted paths, where the weights are determined by the valence of the other terminal vertex of the path. If one adds valence shells at the same distance k for all atoms in a molecule, one obtains the molecular valence shell counts S_k. To make a comparison with the count of paths easier we have divided the shell count by two. Thence s_1 becomes identical to $p_1 + p_2$ and $S_1 = P_1 + P_2$. In Table 2 we listed shell counts s_2, s_3, s_4, ... for the 18 isomers of octane C_8H_{18}. Because isomers have the same number of bonds we see that s_0 is strictly collinear with p_1.

USE OF VALENCE SHELL DESCRIPTORS IN MRA

We have selected three properties of octanes to test the new descriptor: the boiling points (BP), the entropy (S), and the density (ρ). In Table 3 we show the regressions using s_1, s_2, and s_3 in two and three descriptors correlations. For comparison we also show the correlations of the same three properties using the Wiener descriptors W and P (where P = p_3). First, we can observe that the Wiener number and P are describing the correlation of the boiling points of octanes extremely well, the standard error being below 1 °C. Even

Table 3. Regressions Using s_1, s_2, and s_3 in Two and Three Descriptors Correlations for Selected Properties of Octane Isomers[a]

descriptors	s_1, s_2			p_2, p_3
property	*r*	*s*	*F*	
boiling point	0.9188	2.57	41	
entropy	0.9617	1.32	93	
density	0.9915	0.0016	406	
descriptors	**s_1, s_2 s_3**			**p_2, p_3, p_4**
property	*r*	*s*	*F*	
boiling point	0.9721	1.58	80	
entropy	0.9687	1.24	71	
density	0.9932	0.0014	321	
descriptors	**W, P**			**W, P**
property	*r*	*s*	*F*	
boiling point	0.9892	0.96	341	
entropy	0.8860	2.23	27	
density	0.9953	0.0012	732	

[a] The last part gives the correlations using the Wiener descriptors W and P.

better is the regression for the density of octanes using the W and P. On the other hand the two shell descriptors s_1, s_2 describe well the entropy. Use of three shell descriptors improves the correlation of BP considerably and only slightly improved the correlation of entropy. Hence, while W, P still appear to be outstanding descriptors, this is true only for the density and the boiling points of octanes. As we see W and P are not so good descriptors for the entropy of octanes, which is better described by the shells of range two and three.

COMPARISON WITH REGRESSION BASED ON THE PATH NUMBERS

It is of interest to see how well will compare regressions using s_1, s_2 and s_1, s_2, s_3 with regression using p_2, p_3 and p_2, p_3, p_4, respectively. Unexpectedly we found that for all three properties the two sets of descriptors give precisely the same regression coefficient, standard error, and the Fisher ratio. In Figure 2 we show the regression between s_2 and p_3. As we see the two descriptors are not collinear, yet when combined with s_1 and p_2 they give precisely the same correlation statistics, but the corresponding regression equations are different, because the descriptors individually have distinct numerical values. For example, for the boiling points of octane isomers using two descriptors we have the following equations:

$$BP = -4.5529\, p_2 + 2.7615\, p_3 + 133.2905$$

$$BP = -7.3144\, s_1 + 2.7615\, s_2 + 184.4910$$

Both these equations have the same coefficient of regression

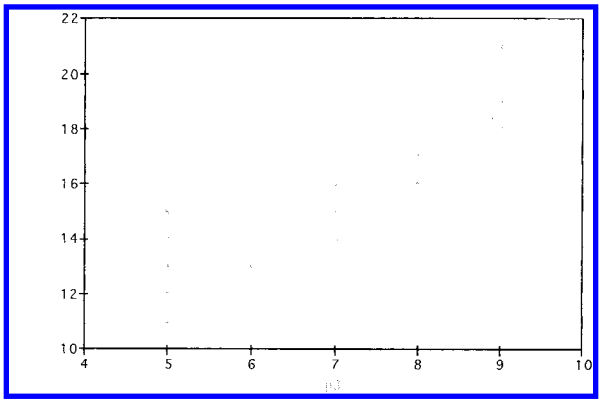

Figure 2. The regressions between p_3 and s_2 for octane isomers.

($r = 0.9188$), the same standard error of estimate ($s = 2.5725$), and the same Fisher ratio ($F = 40.6281$). Also the t-statistics for the corresponding coefficients of the regression equations, which appeared to be isolated in a stepwise regression (i.e., as if descriptors were orthogonalized[10−12]), are the same. For p_2 (and s_1) the t-statistic gives −4.1556, while for p_3 (and s_2) the t-statistic gives 5.5939.

How can we understand the occurrence of the coincidental correlation parameters? If we use p_1 and p_2, or s_0 and s_1, the two characterizations differ only in one descriptor, because p_1 and s_0 are collinear. Completely identical correlation statistics means that the two sets of the descriptors, p_2, p_3 and s_1, s_2, span exactly *the same structure space.* The consequence of this is that s_2 can be expressed as a linear combination of p_1, p_2 . It turns out that for octane isomers not only do we have

$$s_2 = p_1 + p_2$$

$$s_3 = p_2 + p_3$$

but a similar relationship holds for shells and paths of all lengths, i.e.:

$$s_k = p_{k-1} + p_k$$

The above relations, however, do not extend to cyclic structures.

Should s_1, s_2, s_3 be considered distinct novel descriptors in view that they are simple linear transforms of p_1, p_2, p_3 and s_4? While some may view these linear combinations representing the same descriptors, a distinction should be made between simple linear combinations of descriptors such as ($p_1 + p_2$) and use of the same descriptors in linear regression, where particular combinations may arise with a real coefficient as (a p_1 + b p_2) as a result of statistical analysis. For example, Kier and Hall[13] reported a good correlation between (${}^1\chi - {}^1\chi^v$) and the resonance energy for a series of unsaturated and aromatic molecules. That indeed we should view simple linear combinations as distinct descriptors is also evident from the interpretation of such descriptors. Thus while ${}^1\chi$ is an index of molecular branching, the difference (${}^1\chi - {}^1\chi^v$), as pointed out by Kier and Hall, is greater the greater is the degree of unsaturation.

ON LINEAR INDEPENDENCE OF PATHS AND SHELLS

The parallelism between paths and shells may tempt one to discard shells as unnecessary molecular descriptors.

Table 4. A Selection of One and Two Parameter Regressions of Various Properties of Octanes Using Shells, Paths, and the Quotient Paths/Shells as Descriptors

property	descriptor	*r*	*s*	*F*
	Single Descriptor			
MR	s_2	0.8177	0.1088	26.22
steric	s_2	0.9753	0.4056	311.28
	Two Descriptors			
MR	s_1, s_2	0.9960	0.0177	736.89
R^2	s_1, s_2	0.8150	0.1117	14.84
CT	s_1, s_2	0.8538	5.398	20.17
ΔH_f	s_1, s_2	0.9500	0.4123	69.46
S	s_2, s_3	0.9440	1.5874	61.37
ρ	p_3, p_4	0.9752	0.0027	135.69
ρ	s_3, s_4	0.8233	0.0068	14.73
R^2	p_2/s_2, p_3/s_3	0.8512	0.1102	19.72
MR	p_2/s_2, p_3/s_3	0.9922	0.245	379.72

The coincidental statistical parameters of Table 3 may appear as that the valence shell descriptors do not introduce novelty. However, the paths and the shells, except for the already mentioned collinearity of p_2 and s_2, do represent *distinct* descriptors. While it is true that the structure-space spanned by p_2, p_3, p_4 and by s_2, s_3, s_4 is the same space, the subspaces p_3, p_4 and s_3, s_4, are not the same! It is only when these are combined with p_2 (or s_2) that they become identical.

That different pairs of descriptors can span the same space has already been observed by Klein and collaborators.[14] They found that the Wiener index W, or ^{1}W, and the higher order Wiener index ^{2}W as a pair of descriptors span the same space as do the first and the second moments. What is different in our case is that the two sets of the descriptors have one descriptor in common. But what is common to both cases is that in both situations one can characterize the property of interest (here the boiling points, the entropy and the densities of octane isomers) only up to the space defined by any of two *linearly independent descriptors that span the space*. This has a significant consequence for interpretation of regression analyses, as it points to the limitations of interpretation of a regression of molecules by individual descriptors. Instead, as we have just seen, one should place an emphasis on *a set of descriptors* or even not on descriptors but on *the space that descriptors span.* As is well-known different sets of descriptors can span the same space. In view of this the importance of developing suitable bases descriptors, to be used as the reference descriptors, will perhaps become more apparent.

That the paths and the shells do incorporate different structural elements is seen if they are used in simple regressions using a single descriptor. In Table 4 we illustrate a few simple correlations using a single descriptor, merely to point to differences between paths and shells. In the lower part of Table 4 we include several regressions using two descriptors. In addition to s_2, s_3 (and p_2 and p_3) we also considered the quotients p_2/s_2 and p_3/s_3 descriptors, which are reminiscent of path/walk shape descriptors.[15,16] As we see from Table 4 several very respectable regressions for different properties of octanes have been found. It seems therefore that shell descriptors and the quotients of path/shells may find a wider use in structure−property-activity studies. Their apparent advantage is, of course, their very simple and direct structural relationship to valence and extended valence.

REFERENCES AND NOTES

(1) Randić, M. Topological Indices. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; pp 3018−3032.
(2) Platt, J. R. Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys.* **1947**, *15*, 419.
(3) Randić, M.; Razinger, M. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York, 1997.
(4) Randić, M.; Zupan, J. On the interpretation of well-known topological indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550−560.
(5) Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(6) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(7) Randić, M. On the characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(8) CODESSA reports most of the descriptors designed by 1995. Katritzky, A. R., Lobanov, V., Karelson, M. *CODESSA* (*Comprehensive Descriptors for structural and Statistical Analysis*); University of Florida: Gainesville, FL, 1994.
(9) Estrada, E. Novel strategies in the search of topological indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*. Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, The Netherlands, 1999; pp 403−453.
(10) Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 517−525.
(11) Randić, M. Resolution of ambiguities in structure−property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311−370.
(12) Randić, M. Curve fitting paradox. *Int. J. Quantum Chem: Quantum Biol. Symp.* **1994**, *21*, 215−225.
(13) Kier, L. B.; Hall. L. H. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press: New York, 1976; p 231.
(14) Klein, D. J.; Lukovits, I.; Gutman, I. On the definition of the hyper-Wiener index for cycle-containing structures. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 50−52.
(15) Randić, M. Novel Shape Descriptors for Molecular Graphs, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 607−613.
(16) Randić, M. On characterization of shape of molecular graphs. *SAR QSAR Environ. Res.* Submitted for publication.

CI000121I