CROATICA CHEMICA ACTA CCACAA **75** (1) 107–116 (2002)

ISSN-0011-1643
CCA-2785 *Original Scientific Paper*

On the Concept of Molecular Complexity

Milan Randić[a] *and Dejan Plavšić*[b,*]

[a]*Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311, USA and Laboratory of Chemometrics, National Institute of Chemistry, Ljubljana, Hajdrihova 19, Slovenia*

[b]*The Ruđer Bošković Institute, P.O.B. 180, HR-10002 Zagreb, Croatia*

Received October 22, 1998; revised December 13, 2001; accepted December 14, 2001

The previous measures of the complexity of graphs, and thence molecular graphs, have been mainly based on the information content of graphs. We argue here that the two concepts, the information content of a graph and the graph complexity, are distinctive and should be differentiated. We propose a new index of molecular complexity which takes into account not only the connectivity and the closely associated structural features of molecular structure (*e.g.,* branching, cyclicity) but also the symmetry of a molecule as the basis for the partitioning of molecular components considered for construction of the complexity measure of a graph.

Key words: complexity of graphs, molecular graphs, augmented valence complexity index (*AVC*).

INTRODUCTION

The notion of the complexity of molecules has received attention of chemists as can be seen from the recent reviews of the subject.[1,2] The first papers that are explicitly concerned with molecular complexity in chemistry were initiated by Bertz.[3–5] He proposed a measure of the complexity of a molecule, the complexity index $C(n)$, using concepts from information theory and the representation of the molecule by (molecular) graph and its characterization by a graph invariant. The $C(n)$ index is defined by the expression[3]

* Author to whom correspondence should be addressed. (E-mail: dplavsic@rudjer.irb.hr)

$$C(n) = 2n \log_2 n - \sum n_i \log_2 n_i \,, \tag{1}$$

where n denotes a graph invariant and n_i is the cardinal number of the i-th set of equivalent structural elements on which the invariant is defined. The summation goes over all sets of equivalent structural elements. Eq. (1) is a modification of the expression for the information content of a system, I, having N elements[6]

$$I = N \log_2 N - \sum N_i \log_2 N_i \,, \tag{2}$$

where N_i is the cardinal number of the i-th set of elements and the summation runs over all sets of elements.

The aforementioned formulas are based on Shannon's formula[7] for the mean information content of a signal. Rashevsky[8] was the first to apply the Shannon formula to graphs. He calculated the »information content« (per vertex), $\bar{I}$, of a graph,

$$\bar{I} = - \sum p_i \log_2 p_i \,, \tag{3}$$

where $p_i = n_i / n$; n_i is the number of vertices in the i-th set of equivalent vertices, and n is the total number of vertices. Mowshowitz[9] interpreted n_i as the cardinality of orbit i of the automorphism group of a graph and suggested the quantity $\bar{I}$ as a measure of the relative complexity of graphs. Others followed by modifying the »information content« formula and by considering different structural elements to come to alternative measures of the »complexity« of (molecular) graphs.[10,11] Thus Trucco[12] considered the partition of edges of a graph as the basis for the calculation of the information content. Bonchev and Trinajstić[6] put forward information for adjacency, incidence, polynomial coefficients of the adjacency matrix, and for distances of molecular graphs. Bertz[3] derived his formula (Eq. 1) using the number of connections, defined as the number of pairs of adjacent edges in a hydrogen-suppressed molecular graph. Basak and collaborators[13] derived their complexity indices on the basis of the first-order topological neighbourhood of atoms, and recently Hendrickson and co-workers[14] considered the number of hydrogen atoms attached to an atom and the presence of double bonds and triple bonds.

INFORMATION CONTENT *versus* COMPLEXITY

Information content, as defined by the Shannon formula, is well-defined concept. As we have seen all the aforementioned measures of complexity are based on the information content relative to the partitioning of a structural invariant selected for consideration. Each such measure may be of special

interest for a particular consideration, but do they represent complexity? Complexity, just as many other widely used concepts in chemistry (*e.g.*, aromaticity), has not been rigorously defined, at least for a general case. Bonchev and Polansky[15] proposed some desirable attributes for a complexity index, but they are open to discussion. For example, among the requirements is the uniqueness. Even though this may be desirable we feel that this requirement may be viewed as unwarranted, at least when one uses the word complexity in its usual context. It precludes a possibility that two distinct objects have the same complexity. Why should two distinct systems not be equally complex?

We consider here an alternative approach to the molecular and (molecular) graph complexity that differs from the information content of these objects as derived from the Shannon-type formula. It should be mentioned that other kinds of measures of information not related to the information theoretical approach have been put forward. For example, Kolmogorov[16] considered as a measure of information the minimal length of a program that transforms one object into another. The complexity of a graph has also been expressed directly by selected graph invariants, such as the number of paths,[17] the SMM_1 and SMM_2 indices[18] (the symmetry-modified Zagreb M_1 and M_2 indices respectively), Rückers' *twc* (total walk count) index,[19,20] and the number of subgraphs.[20–25] Clearly, whatever approach is taken the complexity remains a relative concept expressing a measure of the level of interrelation of parts of a system that depends on the components selected for examination. However, besides the contribution originating from the size of a system, a measure of complexity has to deal with the symmetry and the complexity of the elements partitioned in each of the equivalence class. We will outline one such measure of complexity and will compare it with the measures of complexity reported for some small graphs by Bertz,[5] Bonchev,[25] and Hendrickson *et al.*[14]

AUGMENTED VERTEX VALENCE

Morgan introduced the notion of extended connectivity[26] in his efforts to arrive at relatively simple canonical scheme for labelling of atoms in a structure. Extended connectivities of a vertex in a graph are obtained by assigning initially to each vertex its connectivity value and then by the iterative process in which at each step previously obtained extended connectivities of the nearest neighbours are added to constitute new extended connectivity of the vertex. The idea is simple, computation straightforward, and despite inherent limitations, Morgan's extended connectivity has received due attention in the chemical literature.[19,27–32] One of the limitations of extended con-

nectivity that has been pointed out in the literature is the occasionally oscillatory behaviour of extended connectivity for selected atoms. Recently, we have considered modification of extended connectivity.[33] Rather than considering a sequence of extended connectivities for a vertex in the alternative approach the valences of neighbouring vertices are added to yield a single numerical value to each vertex. This number is based on the use of different weights for valence contributions of vertices at different distance from the vertex under consideration. The newly constructed valences are referred to as augmented vertex valences or regressive vertex degrees.[33,34]

In Table I we illustrate the construction of augmented valences of the vertices of two smaller bicyclic graphs mentioned in Ref. 5, illustrated in Figure 1. The assumed weights are given by the expression $1/2^d$, where d is

TABLE I

The construction of augmented valences of the vertices of two graphs depicted in Figure l

Vertex	Augmented vertex valence	
	Graph A	Graph B
a	1 + 4/2 + 7/4 + 2/8 = 5	1 + 3/2 + 5/4 + 5/8 = 4.375
b	4 + 8/2 + 2/4 = 8.5	3 + 6/2 + 5/4 = 7.25
c	2 + 7/2 + 5/4 = 6.75	3 + 8/2 + 3/4 = 7.75
d	3 + 8/2 + 3/4 = 7.75	2 + 6/2 + 5/4 + 1/8 = 6.375
e	2 + 5/2 + 6/4 + 1/8 = 6.125	3 + 7/2 + 3/4 + 1/8 = 7.375
f	2 + 6/2 + 6/4 = 6.5	2 + 6/2 + 6/4 = 6.5

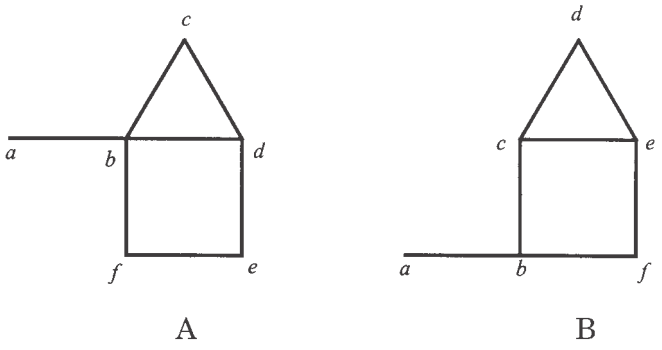

Figure 1. Two labelled bicyclic graphs having the same complexity as reported by Bertz (Ref. 5).

the distance between a neighbouring vertex and the one under consideration. Hence the augmented valence of the vertex labelled with *a* in the first structure (A) is obtained by adding to its valence of one a one half of the valence of the vertex *b*, 1/4 of the sum of valences of the vertices *c*, *d*, and, *f*, and finally 1/8 of the valence of the vertex *e*. If the augmented valences of all vertices of a graph are added up one obtains the graph invariant which we call the augmented valence sum, *AVS*. For the structures A and B of Figure 1 *AVS* is equal to 40.625 and 39.625 respectively.

As one can see from the way *AVS* is constructed the process takes into account the size of a structure indirectly. Equally the *AVS* takes into account the increase in the number of edges (which is a measure of the density of a graph) through the valences of contributing vertices. Hence, at least for graphs showing no symmetry, *i.e.*, for graphs for which a partitioning used as the basis for complexity analysis fully discriminates all vertices, *AVS* appears as a useful measure of graph complexity, which is expected to increase with the size and the density of a graph.

NEW MEASURE OF COMPLEXITY

In our view in order to arrive at a useful measure of complexity of a graph one has to take into account the symmetry of the graph. Presence of symmetry in general reduces the complexity of a system. A way to incorporate the simplifying aspect of symmetry is first to partition the vertex set of a graph into equivalence classes and then to consider the contributions from a single member of each equivalence class only. An argument supporting such an approach rests on the fact that members of an equivalence class are indistinguishable and hence do not contribute new information. This approach however does not keep the count of the number of elements in each equivalence class, a factor that may be of some interest. In more general approach one can introduce a weight for each equivalence class. If the weight equals the number of elements in an equivalence class one obtains *AVS*. One can also choose other weights for individual equivalence classes. This generalization, however, is outside the scope of the present paper.

We put forward a new measure of complexity of a graph (molecule) that we call the *AVC* (augmented valence complexity) index defined as the sum of augmented valences of a single member of each equivalence class of vertices of the graph. To illustrate the *AVC* index let us consider first the nine graphs of Figure 2 that Bonchev examined in Ref. 25. Bonchev calculated the complexity of the graphs by means of his *TC*, *TC*1 and *K* indices. All the three indices give the same order of the nine graphs in respect to increasing complexity (see Figure 2). An intuitive perception of complexity disagrees

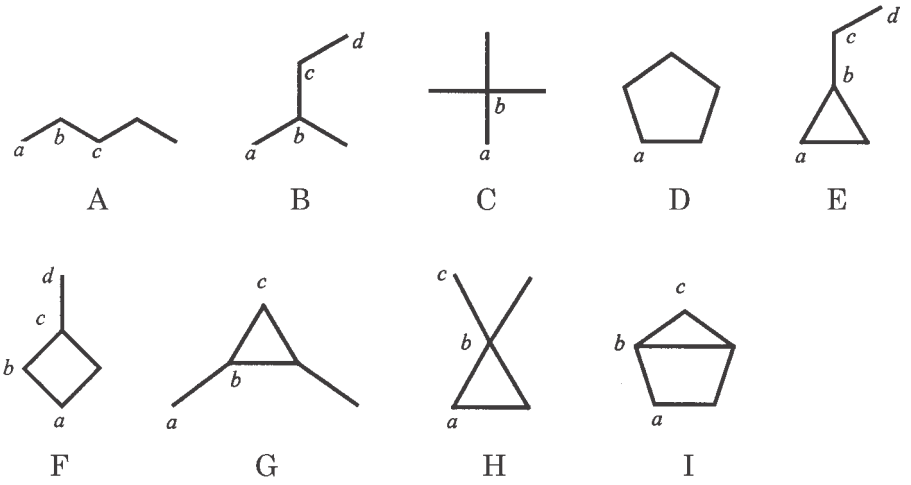

Figure 2. The nine graphs with five vertices whose complexities are given in Tables II and III. For each graph only symmetry non-equivalent vertices are labelled.

with this order of the nine graphs, primarily because Bonchev did not take into account the symmetry of the graphs. Namely, the ring C_5, which has all vertices equivalent, is intuitively expected to qualify as the most simple among the nine graphs, and the graph of neopentane having two kinds of vertices (central and terminal) would be the second most simple structure of Figure 2. Next in the order of increasing complexity one expects graphs having three equivalence classes of vertices and the last the graphs having four different kinds of vertices. The values of the *AVC* index based on the weights $1/2^d$, $AVC(1/2^d)$, for the nine graphs of Figure 2 are shown in Table II. As one can see the new complexity measure $AVC(1/2^d)$, which takes into account symmetry, agrees to a great extent with the common expectations just indicated.

AN ALTERNATIVE COMPLEXITY MEASURE

The $AVC(1/2^d)$ index is but one of the possible complexity measures that incorporate the size, density, and symmetry of a graph. Clearly, instead of the weights given by the expression $1/2^d$ alternative weights can be considered. We will outline here only one alternative scheme for weighting given by the expression $1/d$, where d is the distance between a neighbouring vertex and the one under consideration. By definition for $d = 0$ the weighting factor is equal to zero. In other words, in contrast to the computation of augmented vertex valences in the framework of the weights $1/2^d$ in the alternative scheme

TABLE II

The augmented valences of the symmetry non-equivalent vertices of the nine graphs depicted in Figure 2 and the values of the $AVC(1/2^d)$ index for the graphs

Graph	Vertex	Augmented valence	$AVC(1/2^d)$
A	*a*	1 + 2/2 + 2/4 + 2/8 + 1/16 = 2.813	11.438
	b	2 + 3/2 + 2/4 + 1/8 = 4.125	
	c	2 + 4/2 + 2/4 = 4.5	
B	*a*	1 + 3/2 + 3/4 + 1/8 = 3.375	16.125
	b	3 + 4/2 + 1/4 = 5.25	
	c	2 + 4/2 + 2/4 = 4.5	
	d	1 + 2/2 + 3/4 + 2/8 = 3	
C	*a*	1 + 4/2 + 3/4 = 3.75	9.75
	b	4 + 4/2 = 6	
D	*a*	2 + 4/2 + 4/4 = 5	5
E	*a*	2 + 5/2 + 2/4 + 1/8 = 5.125	19.625
	b	3 + 6/2 + 1/4 = 6.25	
	c	2 + 4/2 + 4/4 = 5	
	d	1 + 2/2 + 3/4 + 4/8 = 3.25	
F	*a*	2 + 4/2 + 3/4 + 1/8 = 4.875	19.875
	b	2 + 5/2 + 3/4 = 5.25	
	c	3 + 5/2 + 2/4 = 6	
	d	1 + 3/2 + 4/4 + 2/8 = 3.75	
G	*a*	1 + 3/2 + 5/4 + 1/8 = 3.875	15.625
	b	3 + 6/2 + 1/4 = 6.25	
	c	2 + 6/2 + 2/4 = 5.5	
H	*a*	2 + 6/2 + 2/4 = 5.5	16.75
	b	4 + 6/2 = 7	
	c	1 + 4/2 + 5/4 = 4.25	
I	*a*	2 + 5/2 + 5/4 = 5.75	18.75
	b	3 + 7/2 + 2/4 = 7	
	c	2 + 6/2 + 4/4 = 6	

for weighting the valence of the vertex under consideration is not included in the construction of its augmented valence. Hence, for example, the augmented valences of the three symmetry non-equivalent vertices *a*, *b*, and *c*

of the graph A of Figure 2 are as follows: 2 + 2/2 + 2/3 + 1/4; 3 + 2/2 + 1/3; and 4 + 2/2 respectively. The difference between the weights $1/2^d$ and $1/d$ is that in the former case the role of more distant neighbours decreases exponentially (simulating the short range influence) and in the latter case the role of neighbours at larger separation decreases at a lesser rate (simulating long range influence). In Table III the values of $AVC(1/2^d)$, $AVC(1/d)$, $AVS(1/2^d)$, and the TC, $TC1$, and K indices for the nine graphs of Figure 2 are listed for comparison. As one can see Bonchev's indices parallel to great extend AVS, the index that does not take into account the symmetry of a graph. Moreover, from Table III one also sees that the complexity measures here introduced discriminate among the small graphs. This was not the case with several of the complexity measures based on the Shannon formula. Thus, for example, the $C(n)$ index based on vertices and edges assigns equal complexities 31.20 and 39.30 respectively to the two graphs of Figure 1. Similarly, the approach of Hendrickson[14] results in over 15 pairs (and occasionally triplets) of smaller graphs (on five or six vertices) having the same complexity. The occur-

TABLE III

The values of the $AVC(1/2^d)$, $AVC(1/d)$, $AVS(1/2^d)$, TC, $TC1$, and K indices for the nine graphs depicted in Figure 2

Graph	$AVC(1/2^d)$	$AVC(1/d)$	$AVS(1/2^d)$	TC	$TC1$	K
A	11.438	13.25	18.376	60	40	15
B	16.125	18.5	19.5	76	55	17
C	9.75	9.5	21	100	64	20
D	5	6	25	160	110	26
E	19.625	23.666	24.75	172	112	27
F	19.875	24	25.125	190	126	28
G	15.625	19.333	25.75	212	136	31
H	16.75	19.5	26.5	230	146	33
I	18.75	23.5	31.5	482	310	54

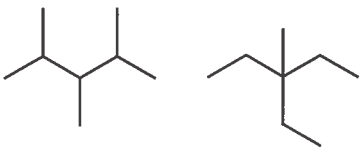

Figure 3. Graphs depicting carbon skeletons of 2,3,4-trimethylpentane and 3-methyl-3-ethylpentane, the two molecules having the same AVC index.

rence of graphs with the same complexity based on augmented vertex valence is, as discussed elsewhere,[33] much less frequent. Figure 3 illustrates one such pair.

Acknowledgment. — This work was supported in part by the Project J-1-8901 of the Ministry of Science and Technology of the Republic of Slovenia and by the Ministry of Science and Technology of the Republic of Croatia. We would like to thank the reviewers for their valuable comments.

REFERENCES

1. L. B. Kier and B. Testa, *Adv. Drug Res.* **26** (1995) 1–43.
2. D. Bonchev and W. A. Seitz, *The Concept of Complexity in Chemistry,* in: D. H. Rouvray (Ed.), *Concepts in Chemistry: Contemporary Challenge,* Research Studies Press, Taunton, U. K., 1996, pp. 353–381.
3. S. H. Bertz, *J. Am. Chem. Soc.* **103** (1981) 3599–3601.
4. S. H. Bertz, *J. Am. Chem. Soc.* **104** (1982) 5801–5803.
5. S. H. Bertz, *Bull. Math. Biol.* **45** (1983) 849–855.
6. D. Bonchev and N. Trinajstić, *J. Chem. Phys.* **67** (1977) 4517–4533.
7. C. E. Shannon, *Bell Syst. Tech. J.* **27** (1948) 379–423.
8. N. Rashevsky, *Bull. Math. Biophys.* **17** (1955) 229–235.
9. A. Mowshowitz, *Bull. Math. Biophys.* **30** (1968) 175–204.
10. D. Bonchev, *MATCH* **7** (1979) 65–112.
11. D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures,* Research Studies Press, Chichester, 1983.
12. E. Trucco, *Bull. Math. Biophys.* **18** (1956) 129–135.
13. A. B. Roy, S. C. Basak, D. K. Harris, and V. R. Magnuson, *Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Activity,* in: X. J. R. Avula, R. E. Kalman, A. I. Liapis, and E. Y. Rodin (Eds.), *Mathematical Modelling on Science and Technology,* Pergamon Press, New York, 1984, pp. 745–750.
14. J. B. Hendrickson, P. Huang, and A. G. Toczko, *J. Chem. Inf. Comput. Sci.* **27** (1987) 63–67.
15. (a) D. Bonchev and O. E. Polansky, *Stud. Phys. Theor. Chem.* **51** (1987) 126–158. (b) D. Bonchev, *The Problems of Computing Molecular Complexity,* in: D. H. Rouvray (Ed.), *Computational Chemical Graph Theory,* Nova Science Publishers, New York, 1990, pp. 33–63.
16. A. N. Ko1mogorov, *Probl. Peredachi Inf.* **1** (1965) 3–11.
17. M. Randić, G. M. Brissey, R. B. Spencer, and C. L. Wilkins, *Comput. Chem.* **3** (1979) 5–13.
18. S. Nikolić, I. M. Tolić, N. Trinajstić, and I. Baučić, *Croat. Chem. Acta* **73** (2000) 909–921.
19. G. Rücker and C. Rücker, *J. Chem. Inf. Comput. Sci.* **33** (1993) 683–695.
20. G. Rücker and C. Rücker, *J. Chem. Inf. Comput. Sci.* **41** (2001) 1457–1462.
21. B. Mohar, *Stud. Phys. Theor. Chem.* **63** (1989) 1–8.
22. E. C. Kirby, R. B. Mallion, and P. Pollak, *Mol. Phys.* **83** (1994) 599–602.
23. S. Nikolić, N. Trinajstić, A. Jurić, and G. Krilov, *Croat. Chem. Acta* **69** (1996) 883–897.

24. S. H. Bertz and T. J. Sommer, *Chem. Commun.* (1997) 2409–2410.
25. D. Bonchev, *SAR & QSAR Environ. Res.* **7** (1997) 23–43.
26. H. L. Morgan, *J. Chem. Doc.* **5** (1965) 107–113.
27. M. Randić, *J. Chem. Inf. Comput. Sci.* **15** (1975) 105–108.
28. M. Razinger, *Theor. Chim. Acta* **61** (1982) 581–586.
29. M. Razinger, *Theor. Chim. Acta* **70** (1986) 365–378.
30. G. Rücker and C. Rücker, *J. Chem. Inf. Comput. Sci.* **31** (1991) 123–126.
31. J. Figueras, *J. Chem. Inf. Comput. Sci.* **33** (1993) 717–718.
32. C. Rücker and G. Rücker, *J. Chem. Inf. Comput. Sci.* **34** (1994) 534–538.
33. M. Randić and D. Plavšić, *Int. J. Quant. Chem.* (submitted)
34. M. V. Diudea, O. Minailiuc, and A. T. Balaban *J. Comput. Chem.* **12** (1991) 527–535.

SAŽETAK

O pojmu složenosti molekula

Milan Randić i Dejan Plavšić

Najveći broj prethodno predloženih mjera složenosti grafa temelji se na količini informacije u grafu. U radu je pokazano da pojam količine informacije u grafu i pojam složenosti grafa nisu istovjetni. Predložen je novi indeks složenosti molekule, koji uzima u obzir ne samo povezanost u molekuli i njoj bliske strukturne odlike (npr. grananje, prstenastost) već i simetriju molekule kao osnovu za razdiobu komponenata molekule na kojima se mjera složenosti temelji.