

## Chapter 4

# Shannon's Information and Complexity

**D. Bonchev**

Department of Marine Sciences, Program for Theory of Complex  
Natural Systems, Texas A&M University, Ford Crockett Campus, 5007  
Avenue U, Galveston, Texas 77551, USA

4.1 Introduction .....	157
4.2 The Notion of Entropy in Physics .....	158
4.3 Information Content of Structures .....	160
4.4 Symmetry-Based Information-Theoretic Indices .....	161
4.5 Magnitude-Based Information Measures of Complexity .....	170
4.6 Substructure-Based Information Measures of Complexity .....	175
4.7 Concluding Remarks .....	180
4.8 References .....	182

## 4.1 Introduction

The intuitive idea of complexity as an antonym of simplicity is an inherent one for the people of the twenty-first century. We live in a complex environment and deal with a complex technology. Every day we face the challenging complexity of life at our work and at home. We enjoy the complexities of our education and entertainment. Could, however, the fuzzy idea of complexity be transformed into a rigorously defined scientific notion? Could one quantitatively assess complexity and, if the answer is “yes,” why is it needed?

In this chapter, we will try to shed some light on the above-mentioned questions insofar as they refer to the realm of chemistry. However, the approach we adopt is a general one and it could be applied to any system having a “structure,” i.e., to any system composed of certain parts or

elements united in a single entity by certain relationships. Processes are not excluded from our consideration provided they include mutually related steps, catalytic chemical reactions being a typical example [1–3].

Consider a couple of chemically relevant cases. Two important classes of industrially produced plastics are high-density polyethylene (HDPE) and low-density polyethylene (LDPE). HDPE is composed mainly of linear macromolecules, whereas in LDPE the macromolecules are highly branched (Figure 4.1a). The two types of polyethylene have very different properties: HDPE has good mechanical characteristics and can be used as construction material, whereas LDPE has low toughness but excellent processability and finds applications for packings. It is intuitively clear that the branched LDPE structures are more complex than the linear HDPE ones. Thus, complexity strongly influences the properties of this material.

Another example is provided by molecules containing the same number of benzene rings connected in a ribbon, a two-dimensional array, and a macro ring (Figure 4.1b). This sequence of benzenoid structures can be extended even further to include three-dimensional shapes, such as spheres (fullerenes), cylinders (nanotubes), etc. These classes of benzenoid structures differ in their chemical reactivity and physical properties. Complexity comparisons between these classes can no longer rely on intuition only; a quantitative measure of complexity is required. Such a relevant measure could be very useful in structure-property relationships, enabling the more effective search for new materials.

## 4.2 The Notion of Entropy in Physics

Clausius introduced entropy in thermodynamics in the mid-nineteenth century to characterize the changes occurring during irreversible processes in systems that do not interact with their surroundings (isolated systems). According to the Second Law of Thermodynamics, such processes (common examples are heat exchange and diffusion) are always associated with an increase in entropy. Half a century later, Boltzmann revealed the statistical character of the Second Law. The only processes allowed to occur spontaneously are those that increase the disorder in systems. Thus, two gases mix irreversibly; they never separate without outside intervention. Similarly, two gases with different temperatures spontaneously equalize their temperature but the process cannot be reversed without some compensation.

A key role in Boltzmann's statistical thermodynamics is played by his famous formula relating the entropy  $S$  to the thermodynamic probability  $W$ ;

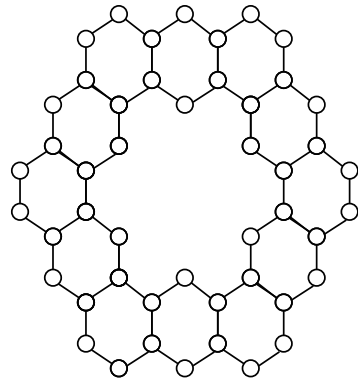
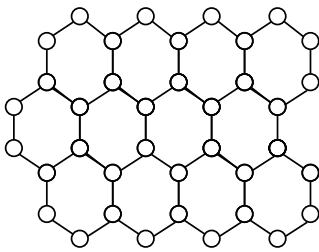
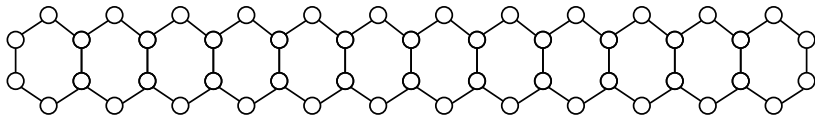
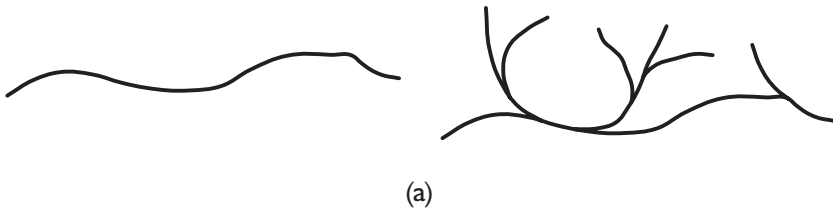


Figure 4.1: (a) Linear and branched macromolecules characterize high-density polyethylene (HDPE) and respectively low-density polyethylene (LPDE); (b) A ribbon, a two-dimensional array, and a macro ring of benzene rings in benzenoid hydrocarbons.

$$S = k \ln W \tag{4.1}$$

where  $k$  is a proportionality constant. Thermodynamic probability is equal to the number of microstates in a system macrostate, characterized by a constant total energy. For a system having  $N$  particles distributed such that  $N_1$  of them have energy  $E_1$ ,  $N_2$  of them have energy  $E_2$ , etc.,

$$W = \frac{N!}{N_1! N_2! \dots N_k!} \tag{4.2}$$

For large  $N$ , the Stirling transformation changes formula (4.2) into

$$\ln W = N \ln N - \sum_{i=1}^k N_i \ln N_i \quad (4.3)$$

### 4.3 Information Content of Structures

In 1949, Shannon [4] showed that the statistical concept of entropy can be extended beyond thermodynamics and applied to the process of transmitting information. Shannon's information theory regards a message transmitted through information channels as a specific set of symbols (an "outcome") selected from the ensemble of all  $k$  such sets containing the same total number of symbols  $N$ . Probabilities  $p_1, p_2, \dots, p_k$  are assigned to each of the outcomes, and the probability of the  $i$ th outcome is proportional to the number of symbols  $N_i$  it contains ( $p_i = N_i / N$ ). Shannon's entropy of information  $H$  characterizes the uncertainty of the expected outcome. When the transmission is totally random all outcomes are equally probably and the entropy of information is the maximal one. Conversely, if there is a single possible outcome,  $H = 0$ . In the intermediate real-life cases the amount of information  $I$  transmitted is the difference between the maximum entropy and the specific value that Shannon's  $H$  function has for the system of interest. Thus, information has the meaning of reduced uncertainty of the final outcome.

The basic Shannon formula is obtained from equations (4.1) and (4.2), after taking the proportionality constant  $k = 1 / \ln 2$  to measure the entropy of information in bits (binary digits).

$$H = N \log_2 N - \sum_{i=1}^k N_i \log_2 N_i, \text{ bits} \quad (4.4)$$

Another form of Shannon's equation determines the average entropy of information per communication symbol:

$$H_{av} = \frac{H}{N} = - \sum_{i=1}^k p_i \log_2 p_i = - \sum_{i=1}^k \frac{N_i}{N} \log_2 \frac{N_i}{N}, \text{ bits/symbol} \quad (4.5)$$

One bit of information,  $H_{av} = 1$  bit, is obtained in learning the outcome of a situation in which there is a choice between two possible options.

One of the major consequences of Shannon's theory was the radically new idea of viewing a *structure* of any kind as a communication, which carries a certain amount of information. Thus, the notion of the *information*

*content of a molecule* emerged in the early 1950s (Dancoff and Quastler [5], Linhitz [6], Rashevsky [7]), along with a generalization of the Second Law Thermodynamics that included information (Brillouin [8]). The Negentropy Principle of Information of Brillouin<sup>8</sup> regards information as a negative component of entropy. In this way, the generalized Second Law allows only such spontaneous processes in isolated systems that increase entropy or/and decrease information. Information cannot increase in irreversible processes; it can only diminish.

Mowshowitz [9] first presented in 1968 a rigorous reinterpretation of Shannon's H-function as information content but not entropy. He pointed out that Shannon's function does not measure the average uncertainty per structure of a given ensemble of all structures having the same number of elements, e.g., the selection of a molecule from the ensemble of all molecules having the same number of atoms. Rather, it is the information content of the structure relative to a system of symmetry transformations that leave the system invariant. In 1979, Bonchev [10] added the argument that entropy is transformed into information by the process of structure formation from its constituent elements. This makes the information a *bonded* one; it is conserved within the system until it is destroyed.

As could be anticipated from the preceding text, this chapter will not represent a general review or comparative analysis of complexity measures used in the realm of chemistry. Such an analysis is presented in the first several chapters of this volume. Here, we focus on the important question whether complexity measures based on Shannon's information theory could at all be adequate measures of *structural* complexity. Serious doubts about the positive answer of this question were raised in our previous publications [11–13]. It will be shown that the positive result is necessarily associated with a reformulation of the original information-theoretic formalism.

#### 4.4 Symmetry-Based Information-Theoretic Indices

The information content of a structure, as defined in Section 4.3, is based on symmetry. The elements of the system are grouped into equivalence classes, according to a certain equivalence criterion or symmetry operation(s) that exchange the elements without violating the structure adjacency. Mowshowitz [9] formalized the application of Shannon's equations to finite systems with symmetry elements. He introduced a probability scheme applicable to any system having  $N$  elements partitioned into  $k$  classes according to the equivalence criterion  $\alpha$  :

Equivalence classes	1, 2, ..., $k$
Element partition	$N_1, N_2, \dots, N_k$
Probability distribution	$p_1, p_2, \dots, p_k$

Here,  $p_i = N_i / N$  is the probability for a randomly chosen element to belong to class  $i$  having  $N_i$  elements, and  $\sum N_i = N$ . Shannon's equations (4.4) and (4.5) can now be rewritten as equations for the information content  ${}^e I(\alpha)$  of the system, and for the average information content  ${}^e I_{av}(\alpha)$  of a system element. The left superscript  $e$  stands for "equivalence" to distinguish this type of information measure from those for "magnitude" introduced in Section 4.5.

$${}^e I(\alpha) = N \log_2 N - \sum_{i=1}^k N_i \log_2 N_i \quad (4.4a)$$

$${}^e I_{av}(\alpha) = \frac{{}^e I(\alpha)}{N} = -\sum_{i=1}^k p_i \log_2 p_i = -\sum_{i=1}^k \frac{N_i}{N} \log_2 \frac{N_i}{N} \quad (4.5a)$$

The values of both information indices thus introduced vary within the following ranges:

$$0 \leq {}^e I(\alpha) \leq N; \quad 0 \leq {}^e I_{av}(\alpha) \leq 1 \quad (4.6)$$

The upper bound of these ranges is reached when each element forms a separate class,  $N_i = 1$ , i.e., when the system has no symmetry. The information content is zero when all of its elements are equivalent,  $N_1 = N$ , i.e., when the system has no structure, due to its very high symmetry. One might infer that the information indices  ${}^e I(\alpha)$  and  ${}^e I_{av}(\alpha)$  could be used as complexity measures, which relate higher complexity to asymmetry and a larger diversity of system elements. Low complexity (simplicity) is characterized as uniformity or lack of diversity, resulting from high symmetry.

#### 4.4.1 Atomic Information Content

Interesting results are obtained when information theory is applied to atoms and molecules on a quantum-mechanical level. When electrons are regarded as distributed over spin orbitals, according to the Pauli exclusion principle each electron must occupy a different orbital. Thus, all  $N_i$ 's in equations (4.4a) and (4.5a) are equal to 1, and the information content of

the respective atom or molecule is the maximum one for the given number of electrons. The Pauli principle was reinterpreted as a principle for the maximum information content of atoms and molecules or, more generally, of fermionic systems [14]. Conversely, photons or, more generally, bosonic particles can populate the same quantum state in unlimited numbers. All particles in this kind of system are equivalent, and the information content is zero.

Other information-theoretic indices have been introduced for atoms and molecules to distinguish those of them having the same number of particles. *Atomic information content* has been introduced in various ways depending on the distribution of electrons, protons and neutrons in the electron shell and atomic nucleus [15,16]. As an example, consider the electron distribution into atomic orbitals (AOs) for the sulfur atom. According to the Pauli exclusion principle, only one or two electrons can populate an atomic orbital. The sixteen sulfur electrons occupy seven AOs with two electrons each, and two AOs with one electron each. Applying equation (4.4a), one obtains  $I(S, AO) = 16 \log_2 16 - 7.2 \log_2 2 - 2.1 \log_2 1 = 66$  bits per atom, and  $I_{av}(S) = 66/16 = 4.125$  bits per electron.

In assessing the complexity of atoms, information-theoretic indices have found application by providing a new systematic of nuclides [15], a reinterpretation of the Periodic Table of the chemical elements [17], and predictions of the properties of transactinide elements 113–118 [18]. A trend has been found which shows that the atomic information content increases the most when a new electron shell or s-, p-, d- and f-subshell begins. This made it possible to predict the atomic number of the first g-element [17].

#### 4.4.2 Molecular Information Content

The information-theoretic formulas (4.4a) and (4.5a) have been used to characterize molecular complexity in various ways, depending on the structural elements taken as a basis, and the equivalence criterion used to group these elements into classes. The most straightforward way is to proceed from the chemical formula of the compound, which represents the elements incorporated as well as the ratio in which the atoms of the elements occur. Thus, Dancoff and Kastler [5] defined the information on the chemical composition  ${}^e I(cc)$  of a molecule in 1953 by using the nature of the chemical element as an equivalence criterion. As an example, consider the molecule of potassium hydrogen phosphate,  $K_2HPO_4$ . The eight atoms of this molecule are distributed into four elemental classes, and the formulas (4.4Aa) and (4.5a) produce  ${}^e I(cc) = 8 \log_2 8 - 4 \log_2 4 - 2 \log_2 2 - 2.1$

$\log_2 1 = 14$  bits per molecule and  ${}^e I_{av}(cc) = 14/8 = 1.75$  bits per atom. Information on the chemical composition of a molecule demonstrates the potential of the information content of a system to characterize its complexity regarded as the diversity of its elements. An illustration of this conclusion is the series of halogen derivatives of methane.  ${}^e I(cc)$  increases systematically in the sequence of molecules  $\text{CH}_4$ ,  $\text{CH}_3\text{F}$ ,  $\text{CH}_2\text{FC1}$ ,  $\text{CHFC1Br}$ : 3.61, 6.86, 9.61, and 11.61 bits per molecule; respectively.

Unlike information on elemental composition, *structural information content* can be characterized in a variety of ways, proceeding from the molecular geometry or from the connectivity of atoms. Regarding atoms as the simplest structural elements of a molecule, one may use symmetry to group them into equivalence classes. When symmetry is characterized by the corresponding point group of the molecule one defines the *information index on molecular symmetry*,<sup>19</sup>  ${}^e I_{sym}$ .

Alternatively, when only atom-atom connectedness is taken into consideration but not specific molecular geometry, a molecule is represented by a molecular graph. The simplest criterion for equivalence of the graph vertices is their degree, i.e., the number of their nearest neighbors, and this structural information index is termed the *information on the vertex degrees*,  ${}^e I(deg)$ . Despite its simplicity this criterion is very relevant in chemistry where it corresponds to the coordination number in crystals and coordination compounds. It provides the classification of carbon atoms in organic chemistry as primary, secondary, tertiary, and quaternary ones.

When the equivalence criterion is extended over all neighboring atoms, the resulting vertex distribution determines the *topological information index* of Rashevsky [7]. The most precise definition of vertex equivalence was given by Trucco [20] and is based on the automorphism group of the graph. Equivalent graph vertices are those that belong to the same orbit of this group of symmetry. Due to the fact that many other topological information indices have been proposed after Rashevsky, we prefer to call this index the *information on the vertex orbits*,  ${}^e I(orb)$  [21].

As an example for the above-mentioned types of structural information content, consider the molecule of 2,2,3,3-tetramethyl pentane and its graph (Figure 4.2). Let the molecule be taken in its most symmetric conformation belonging to the  $C_s$  group. Atoms 1–5 are positioned in the horizontal  $xy$ -plane, whereas the pairs of symmetric atoms 6,7 and 8,9 are in the vertical  $xz$ -plane. The set of nine carbon atoms are then distributed into two classes of two atoms and five classes of one atom each. Equations (4.4a) and (4.5a) produce for this case  ${}^e I_{sym} = 24.53$  bits per molecule and 3.17 bits per atom. The distribution of the nine vertices in the molecular graph

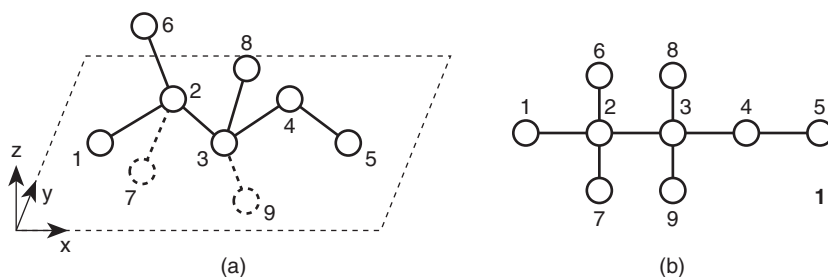


Figure 4.2: (a) The point group symmetry of the 2,2,3,3-tetramethylpentane molecule makes equivalent the pairs of atoms 6,7 and 8,9; (b) The automorphism group of the molecular graph makes equivalent vertices 5,6,7 and 8,9.

according to their vertex degrees is two vertices of degree four, one vertex of degree two, and six vertices of degree one. The resulting value of the information on the vertex degrees is  ${}^e I(deg) = 11.02$  bits per molecule and  ${}^e I_{av}(deg) = 1.22$  bits per atom. When graph vertices are grouped into the graph orbits, vertices 1, 6, and 7 belong to the same orbit. Another orbit includes vertices 8 and 9. The remaining four vertices are in an individual orbit each. The distribution  $9(3,2,1,1,1,1)$  determines the total information on the graph orbits of this molecule to be  ${}^e I(orb) = 21.77$  bits, and the average information is  ${}^e I_{av}(orb) = 2.42$  bits per atom.

Up to here we have introduced three different criteria for determining equivalent atoms in a molecule by regarding atoms as the simplest structural elements. Graph edges can also be used as structural elements and the orbits of the corresponding edge group of the graph can determine their equivalence classes. Trucco mentioned this kind of information index [20]. We shall call it the *information on the edge orbits* of the graph,  ${}^e I(edge\ orb)$ . There are eight edges in the molecular graph **1** of 2,2,3,3-tetramethylpentane (Figure 4.2). According to their symmetry they form five edge orbits:  $\{12, 26, 27\}$ ,  $\{38, 39\}$ ,  $\{23\}$ ,  $\{34\}$ , and  $\{45\}$ . The respective values of the information indices are  ${}^e I(edge\ orb) = 8 \log_2 8 - 3 \log_2 3 - 2 \log_2 2 = 17.25$  bits per molecule, and  ${}^e I_{av}(edge\ orb) = 2.16$  bits per edge.

The next more complex type of substructure after vertices and edges are the subgraphs containing two adjacent edges. Gordon and Kennedy [22] were the first to use this number as a branching index. Even earlier, Platt [23] introduced a twice-larger index as the sum of the first bond neighbors of each bond in the molecule. Bertz defined this type of substructure more broadly as having a “pair of adjacent edges (connections)”

and including multiple edges and loops [24]. One can also define the automorphism group of graph connections. Its orbits determine the *information on connections orbits*,  ${}^eI(\text{conn orb})$ . There are twelve connections in graph **1**, and they are distributed into the following orbits: {126, 127, 267}, {123, 623, 723}, {238, 239}, {438, 439}, {234}, and {345}. The distribution 12 {3,3,2,1,1} yields  ${}^eI(\text{conn orb}) = 29.51$  bits per molecule and  ${}^eI_{av}(\text{conn orb}) = 2.46$  bits per connection.

One might continue applying the Shannon equations to larger and larger substructures. Instead, we will inspect several more indices of this class based on different types of structural invariants. Two such indices have been introduced by proceeding from the equivalence of distances in a molecular graph,  ${}^eI(\text{dist})$  [25,26], and the equivalence of vertices being equidistant to the graph center,  ${}^eI(\text{centric})$  [27,28]. The distance between two vertices is the number of edges along the shortest path that connects them; therefore, all graph distances are integers. Consider again as an example graph **1** (Figure 4.2b). There are 36 distances in this graph: 8 distances of length 1, 13 distances of length 2, 12 distances of length 3, and 3 distances of length 4. The distance distribution 36 {8, 13, 12, 3} results in  ${}^eI(\text{dist}) = 66.24$  bits per molecule and  ${}^eI_{av}(\text{dist}) = 1.84$  bits per distance.

The *centric information index*,  ${}^eI(\text{centric})$  [27,28] is based on the definition of the graph center [29]. The classical definition of Jordan, given in the nineteenth century, specifies the center of trees (acyclic graphs) with a procedure of consecutive pruning the tree by cutting off all of its terminal vertices. The final result is a single center (a vertex) or a bicenter (an edge). In 1969, Harary [26] proposed a rigorous definition that also includes cyclic graphs. It introduced the notion of the *vertex eccentricity*  $e(i)$  as the longest distance from a given vertex to any other vertex in the graph. The graph center is the vertex with the minimum eccentricity:

$$e(i) = \max d(ij) \quad ; \quad e(i) = \min \quad (4.7)$$

The pitfall of this definition is that it often produces a group of nonequivalent central vertices. Bonehev, Balaban and Mekenyan [27] proposed a more detailed solution in which the centric vertices belong to the same orbit of the automorphism group of the graph. The classical definition (4.7) is regarded only as a first of several hierarchically applied criteria, which gradually reduce the number of central vertices until only a single vertex or two or more equivalent vertices remain. The second criterion requires the central vertex to have the smallest sum of distances

to all the remaining vertices in the graph. This sum is termed the *vertex distance*,  $d(i) = \sum_j d(i,j)$ , or *distasum* [30]. If two or more nonequivalent vertices have the same minimal eccentricity and the same vertex distance then the third criterion, requiring a minimal occurrence of the largest distance,  $n(i,j;\max)$ , is used:

$$d(i) = \sum_j d(i,j) = \min \quad (4.8)$$

$$n(i,j;\max) = \min \quad (4.9)$$

When the three criteria fail to produce equivalent centric vertices, an Iterative Vertex-Edge Centricity (IVEC) algorithm is applied [31], which always solves the problem with only one or two iterations.

In the example of graph **1** the first criterion suffices to determine vertex 3 as the graph center, due to its minimal eccentricity  $e(3) = 2 = \min$ . Vertices 2, 4, 8, and 9 have eccentricity 3, and vertices 1, 5, 6, and 7 have eccentricity 4. Let us now order the graph vertices centrically, according to their distance from the graph center. Two layers of equidistant vertices are thus formed around the center with vertices 2, 4, 8, and 9 in the first neighborhood layer, and vertices 1, 5, 6, and 7 in the second one. Thus, the nine vertices of this graph have a centric distribution 9 {1, 4, 4}, from which  ${}^eI(\text{centric}) = 12.53$  bits per molecule and  ${}^eI_{av}(\text{centric}) = 1.39$  bits per atom.

Several information indices have been developed by proceeding from combined equivalence criteria. In fact, the first topological information index of Rashevsky [7] combines two criteria of atom equivalence: in order to be equivalent two atoms in a molecule should belong to the same chemical element and have the same atomic neighborhoods up to the terminal atoms. Basak et al. [32–34] used the same criteria but with separate terms for the first, second, ...,  $k$  th neighborhood. Basak's  $k$ th order *neighborhood complexity index* in hydrogen-suppressed graphs is thus identical to Rashevsky's index, and for such graphs having no heteroatoms his first order index coincides with the information on vertex degrees  ${}^eI(\text{deg})$ . Basak applied his neighborhood indices not only to hydrogen-suppressed molecular graphs but also to molecular graphs that include hydrogen atoms and heteroatoms, and his methodology was used with success in a wide range of QSAR studies and assessments of environmental toxicities [34].

Other information indices based on a combination of structural criteria have also been introduced [10,21,26,35]. Like the indices described above (Table 4.1, *vide infra*), they reflect only some of the complexity features

but fail for others. For this reason, we add here only the *molecular complexity* index of Bertz [24, 36],  $BI$ , which satisfies most of the requirements for a complexity measure. Bertz's index combines the information on graph connections with a size term  $N \log_2 N$ , with  $N$  being the total number of connections:

$$BI = 2N \log_2 N - \sum_{i=1}^k N_i \log N_i \quad (4.10)$$

Adding the size term reduces the inherent pitfalls of symmetry-based information complexity measures. The corrective term allows us to obtain different nonzero values for highly symmetrical molecules like the cycloalkanes. It reflects to a certain extent molecular complexity features like branching and cyclicity. Thus, it increases the structural component of the information measure and diminishes the simplifying effect of symmetry.

It is essential to verify the extent to which the information-theoretic indices introduced so far, proceeding from the differently defined equivalence classes of structural invariants, satisfy the intuitive idea of complexity. Experts generally agree that path (or linear) graphs  $P_n$  are the simplest connected graphs and the complete graphs  $K_n$  are the most complex ones. There is also general agreement that for the same number of vertices the star graphs  $S_n$  and monocyclic graphs  $C_n$  are more complex than path graphs and less complex than complete graphs. Bertz and Zamfirescu [37] formalized these criteria in the form of inequalities which a complexity measure must satisfy.

We extend this series of four graphs with two more, the first being totally disconnected, and the second being bicyclic. Indeed, the intuitive expectation is that the disconnected graph will be less complex than any connected graph including  $P_n$ . Similarly, the bicyclic graph should be more complex than the monocyclic one. Figure 4.3 shows the selected six graphs each having five vertices. The values of the calculated information indices are shown in Table 4.1. Since the first four indices are calculated from the

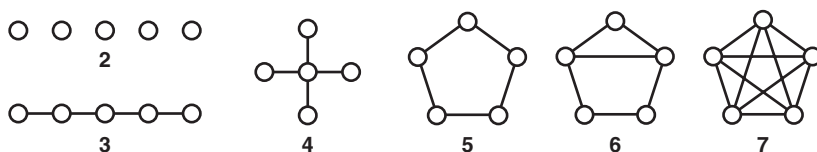


Figure 4.3: Graphs with five vertices used as an illustration of increasing structural complexity.

Table 4.1: Information indices based on the equivalence of vertices, edges, connections, centrally ordered vertices, and distances cannot reproduce the increasing structural complexity of the sequence of the five-vertex graphs **2–7**

Index	Graphs					
	2	3	4	5	6	7
$^eI(\text{vertex}, \text{deg})$	0	7.61	3.61	0	4.86	0
$^eI(\text{vertex})$	0	7.61	3.61	0	7.61	0
$^eI(\text{edge})$	0	4	0	0	11.51	0
$^eI(\text{connection})$	0	2.75	0	0	20.53	0
$^eI(\text{vertex}, \text{centric})$	0	7.61	3.61	0	7.61	0
$^eI(\text{distance})$	0	18.47	9.71	10	9.71	0
BI (Bertz index)	0	2.75	15.51	11.61	49.06	147.21

orbits of the respective automorphism groups, for brevity we omit the abbreviation “orb”, and denote only the type of structural element used: vertices, edges, connections, and centrally ordered vertices. The Basak set of indices [32–34] were not included because for hydrogen-suppressed graphs that have no heteroatoms they include  $I(\text{deg})$  and  $I(\text{vertex orbits})$ .

Although the six graphs have the same number of vertices, the number of edges increases from zero in **2**; to four in **3** and **4**, to five in **5**, to six in **6**, and to ten in **7**. The number of connections increases even faster: from zero in **2**, to three in **3**, to six in **4**, to five in **5**, to nine in **6**, and to twenty one in **7**. Intuition implies that the structure is more complex when the number of the interconnections of its elements increases. However, the first five indices shown in Table 4.1 fail to distinguish between the disconnected graph **2**, the monocycle **5**, and the complete graph **7**; ascribing zero information content to all of them. The information on the equivalence of distances decreases with increasing branching and cyclicity. This is well reflected by the values of star versus path graph, as well as in the sequence monocyclic  $\Rightarrow$  bicyclic  $\Rightarrow$  complete graph. However, the equal complexity of disconnected and complete graphs is a major failure of this index. Another problem of most equivalence-based indices is their opposing complexity trends to decrease with branching and cyclicity and increase with size.

One might conclude that replacement of the vertices, which are the simplest subgraphs, with larger subgraphs containing two and respectively three vertices, as well as the use of graph distances will not produce an adequate complexity measure. The real meaning of these symmetry-based indices is that they measure diversity. Diversity coincides with complexity

when dealing with the chemical composition of molecules. However, as manifested by the examples in Table 4.1, structural diversity and structural complexity have little in common. Yet, the last line in Table 4.1 shows that with a conveniently selected corrective term, the equivalence of some more sensitive graph invariants, like the two-edge subgraphs used by Bertz, could provide better complexity measures. Yet, as can be inferred from the more detailed evaluation [38], the Bertz' index [21] faced difficulties when dealing with more subtle structural patterns, and the search for new solutions continued.

#### 4.5 Magnitude-Based Information Measures of Complexity

An extension of the Shannon information-theoretic approach to the description of chemical structures was reported in 1977 by Bonchev and Trinajstić [25]. The finite probability scheme of Mowshowitz [8] was expanded so as to include certain weights or magnitudes of the structure elements. The scheme can be applied to any system having  $N$  elements partitioned into  $k$  classes according to the element weight (or magnitude) of type  $\alpha$  :

Equivalence classes	$1, 2, \dots, k$
Element partition	$N_1, N_2, \dots, N_k$
Probability distribution	$p_1, p_2, \dots, p_k$
Magnitudes (weights)	$w_1, w_2, \dots, w_k$
Probability M-distribution	${}^m p_1, {}^m p_2, \dots, {}^m p_k$

Here,  $\sum N_i w_i = M$ , where  $M$  is the magnitude of the criterion (or property) selected to partition the system elements,  $p_i = w_i / M$  is the probability for a randomly chosen element to belong to class  $i$  having magnitude  $w_i$  and  $\sum {}^m p_i = 1$ . Shannon's equations (4.4) and (4.5) define in this approach the magnitude-based information content  ${}^m I(\alpha)$  of the system, and the corresponding average information content  ${}^m I_{av}(\alpha)$  per system element:

$${}^m I(\alpha) = M \log_2 M - \sum_{i=1}^k N_i w_i \log_2 w_i \quad (4.11)$$

$${}^m I_{av}(\alpha) = -\sum_{i=1}^k N_i \frac{w_i}{M} \log_2 \frac{w_i}{M} \tag{4.12}$$

Information measures (4.11) and (4.12) are defined within the ranges

$$0 \leq {}^m I(\alpha) \leq M \log_2 M : 0 \leq {}^m I_{av}(\alpha) \leq \log_2 M \tag{4.13}$$

where the lower bound corresponds to a system without a structure ( $w_i = M$ ), and the upper bound can be attained by a system having a maximum number of classes with a single element of unit weight in each class ( $k = N = M$ ).

The magnitude-based information indices were first introduced by Benchev and Trinajstić [25] for the distribution of distances in a molecular graph,  ${}^m I(\text{distance})$  or  ${}^m I_d$ . The total magnitude characterizing the structure here is the *Wiener number* [39]  $W$ , which is the sum of all graph distances (the total graph distance). The classes of distance are those of distances  $d(i) = 1, 2, 3, \dots, d(\text{max})$ , and the number of distances in these classes is  $N_1, N_2, \dots, N_k$ , respectively. Thus the general formulas (4.11) and (4.12) are transformed into

$${}^m I(\text{dist}) = W \log_2 W - \sum_{i=1}^k N_i d(i) \log_2 d(i) \tag{4.14}$$

$${}^m I_{av}(\text{dist}) = -\sum_{i=1}^k N_i \frac{d(i)}{W} \log_2 \frac{d(i)}{W} \tag{4.15}$$

A second magnitude-based information measure is defined from the decomposition of the sum of distances in the graph distance matrix (doubled Wiener number) into contributions of different vertices,  $d_i$ . Termed the *distance degree* of vertex  $i$  (or *distasums*), these quantities represent the sum of the distances between the vertex and the remaining graph vertices. Thus, the vertex decomposition of the total graph distance is presented by the formulas:

$${}^m I(\text{dist deg}) = 2W \log_2 2W - \sum_{i=1}^N d_i \log_2 d_i \tag{4.16}$$

$${}^m I_{av}(\text{dist deg}) = -\sum_{i=1}^N \frac{d_i}{2W} \log_2 \frac{d_i}{2W} \tag{4.17}$$

Graph vertices can be partitioned into magnitude-based classes in various ways. Besides the distance degrees distribution introduced in the previous paragraph, vertex distributions based on other criteria (or “degrees”) can be used. In fact, each symmetry-based distribution can be transformed into a magnitude-based one. The first case of vertex distribution discussed in Section 4.4 dealt with the equivalence of vertex degrees, as inferred from the adjacency matrix of the graph. The sum of the vertex degrees,  $a_i$ ,  $A = \sum a_i$ , is called the graph *total adjacency* [35]. This is the simplest magnitude-type graph invariant that can be used to construct an information descriptor. The vertex degree distribution  $A \{a_1, a_2, \dots, a_k\}$  produces the information index on vertex degree magnitudes [35],  ${}^mI(\text{vert deg})$ :

$${}^mI(\text{vert deg}) = A \log_2 A - \sum_{i=1}^N a_i \log_2 a_i \quad (4.18)$$

$${}^mI_{av}(\text{vert deg}) = - \sum_{i=1}^N \frac{a_i}{A} \log_2 \frac{a_i}{A} \quad (4.19)$$

The graph connections (subgraphs having three connected vertices) used as the basis of the Bertz [24] index  $BI$ , can also be used to construct a vertex distribution  $C \{C_1, C_2, \dots, C_n\}$ , according to the number of connections  $C_i$  beginning in each vertex. Here  $C$  is the doubled number of graph connections. The magnitude-based information index on the graph connections,  ${}^mI(\text{conn deg})$ , we propose here is thus defined as

$${}^mI(\text{conn deg}) = C \log_2 C - \sum_{i=1}^N C_i \log_2 C_i \quad (4.20)$$

$${}^mI_{av}(\text{conn deg}) = - \sum_{i=1}^N \frac{C_i}{C} \log_2 \frac{C_i}{C} \quad (4.21)$$

In Table 4.2 we analyze the capability of the four magnitude-based information indices, discussed in this section, to be used as complexity measures. Several conclusions can be drawn. It is seen that the total information indices show a considerably more consistent complexity trend than the average ones. The two distance-based measures demonstrate a very regular pattern of *decreasing* with the increase in graph complexity. Therefore, these two indices are convenient complexity measures for isomeric molecules. The opposing trend to *increase* with the system size, however, prevents their use as complexity estimates of structures having

Table 4.2: Values of four magnitude-based information indices for graphs 2–7

Index	Graphs					
	2	3	4	5	6	7
${}^mI(\text{distance})$	0	64.93	52.00	48.60	45.30	33.22
${}^mI_{\text{av}}(\text{distance})$	0	3.245	3.250	3.240	3.236	3.322
${}^mI(\text{dist deg})$	0	91.63	73.39	69.66	64.86	46.44
${}^mI_{\text{av}}(\text{dist deg})$	0	2.298	2.293	2.322	2.316	2.322
${}^mI(\text{vert deg})$	0	18.00	16.00	23.22	27.51	46.44
${}^mI_{\text{av}}(\text{vert deg})$	0	2.250	2.000	2.322	2.293	2.322
${}^mI(\text{conn deg})$	0	13.51	24.00	23.22	64.60	139.32
${}^mI_{\text{av}}(\text{conn deg})$	0	2.250	2.000	2.322	2.936	2.322

different size. Nevertheless, these indices have a good potential for QSPR/QSAR applications when used jointly with a convenient size term. The index on the vertex degrees distribution increases with both size and cyclicity [40, 41] (cyclic complexity) but fails to show the same trend with branching [25, 42–45] (acyclic complexity). Yet, the correct trend shown in cyclic molecules indicates that the magnitude-based information indices could in principle provide a good complexity measure if a more detailed structural invariant was taken as a basis.

We investigated this idea by introducing the new  ${}^mI(\text{conn deg})$  index, based on the distribution of graph connections over the vertices from which

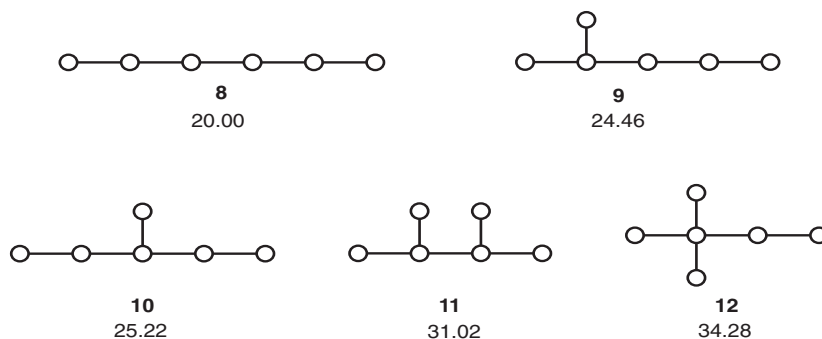


Figure 4.4: Graphs of the five acyclic hexane molecules ordered according to their increasing complexity, as assessed by the magnitude-based information index on the vertex connection degrees,  ${}^mI(\text{conn deg})$ .

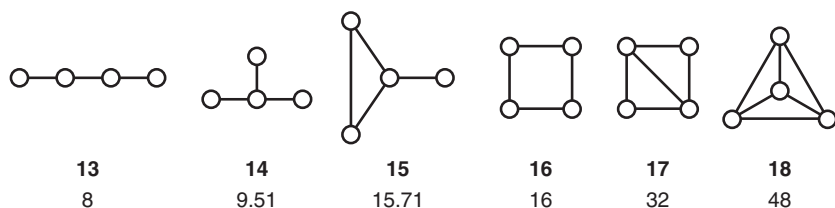


Figure 4.5: All four-vertex graphs ordered according to their increasing complexity.

they emanate. As seen in Table 4.2, the idea is a promising one, the new index showing a very regular increase with the increase in both branching and cyclicity. It also increases with the size of the system as another complexity component. Thus, its values for C3–C6 normal alkanes are 2, 8, 13.51, and 20 bits per molecule, and for C3–C6 cycloalkanes they are 9.51, 16, 23.22, and 31.02, respectively. The sensitivity toward subtler branching patterns can be illustrated with the series of five isomeric acyclic hexanes **8–12** (Figure 4.4). The new information index increases strongly with factors such as the number of branches and the degree of the branched vertex. A moderate increase is found when the branch is shifted toward a more central position.

A similar, systematic increase in the  ${}^mI$  (conn deg) index values with the increase in branching and cyclicity can be seen in Figure 4.5 for all graphs having four vertices. Graphs **13–18** in Figure 4.5 are identical with graphs **VI–XI** in Chapter 2, used there as a basis for comparison of a large set of complexity measures. The ordering provided by our new index: **VI**  $\Rightarrow$  **VII**  $\Rightarrow$  **VIII**  $\Rightarrow$  **IX**  $\Rightarrow$  **X**  $\Rightarrow$  **XI** does not coincide with any of the orderings analyzed in that chapter. It is close to the ordering provided by the Minoli index [47] and the number of spanning (or maximal) trees [3, 48–51], which, however, cannot distinguish between isomeric trees (acyclic graphs).

Thus, the  ${}^mI$  (conn deg) index mirrors the complexity trends in both acyclic and cyclic molecules by proceeding from a general theoretical scheme but without adding a corrective term to the symmetry-based index, as is done in Bertz' treatment of graph connections distributions. More details on the new index will be given in a forthcoming publication [46]. We may conclude that information-theoretic indices based on magnitude distributions provide a better basis for the complexity assessments of molecules than the symmetry-based measures. However, only those magnitude-based distributions of graph invariants that describe the structure in more detail could be good measures of structural complexity. Graph connections

are only the lowest level of a more complete description of a structure. One could easily infer that in graphs whose vertices all have the same degree (e.g., the graphs of the cubane molecule,  $(\text{CH})_8$  and its isomers) one would need to up one level and deal with distributions of three-edge substructures. One might thus also suggest that it is not the specific mathematical function used (Shannon's entropy function) but rather the detailed description of the structure that enables the construction of more reliable complexity measures. Developments along these lines are discussed in the next section.

#### 4.6 Substructure-Based Information Measures of Complexity

A breakthrough in the methods of assessing the complexity of structures has been achieved during the last five years. The novel concept of complexity may be summarized by the sentence: "The more substructures in a system, the more complex the system." This concept is in agreement with our intuitive understanding that while the size of a system contributes to its complexity, it is the connectedness of the system elements that matters more. A larger system with weak interrelations of its elements may be regarded less complex than a smaller system with a high degree of internal connectedness. The more connected the system, the higher the number of substructures  $K$  in it. Then, why not simply count how many substructure there are in a structure?

Bertz and Herndon [52] briefly mentioned such an idea in 1986 in work devoted to measures of molecular similarity. In 1996–1998, simultaneously and independently, Bertz [53, 54] and Bonehev [55–58] developed this complexity concept in detail. Bone and Villar [59] used a similar approach to characterize molecular diversity in detail. Bertz also proposed to use the number of kinds of subgraph  $N_s$  as a measure of structural diversity and applied his complexity measures to determine the "strategic" bond, the creation of which in a synthetic reaction would increase the complexity of the molecule the most. Bonehev developed a substructure-based complexity concept by constructing a *complexity vector*  $K'$

$$K' \{ {}^0K, {}^1K, {}^2K, \dots, {}^EK \}; \quad K(G) = \sum_{e=0}^E {}^eK \quad (4.22)$$

This is an ordered sequence, the  $e$ th *order complexity* term,  ${}^eK$ , in which counts the subgraphs having  $e$  edges;  $E$  in equation (4.22) is the total number of graph edges.

In parallel with the idea of using all substructures, Rucker and Rucker [60] proposed to use *all walks* in the graph. They have shown that *the total walk count twc* is a measure of graph complexity that mirrors the basic patterns of acyclic and cyclic complexity to a very high degree [38].

The substructure approach to complexity was developed further by the present author into the concept of *overall complexity indices* [61–63]. The idea is to ascribe a certain weight to each subgraph, select simple graph invariants as weights, and then to find the overall value of this index for the entire structure by summing up over all subgraphs. Proceeding from vertex degrees and vertex-vertex distances, the graph invariants selected for each subgraph  $i$  were the total adjacency  $A_i$  (the sum of the vertex degrees  $a_i$ ) the total distance  $W_i$  (the Wiener number), and the first and second Zagreb indices [64, 65],  $M1_i$  and  $M2_i$  (the sum of all the squared vertex degrees  $\sum a_i^2$  and the sum of the products of the vertex degrees,  $\sum a_i a_j$ , over all edges  $\{ij\}$ , respectively). The complexity indices thus defined were termed the *overall connectivity* [58, 61], OC, the *overall Wiener number* [62], *OW* and the *overall Zagreb indices* [63], *OM1* and *OM2*, respectively. The overall connectivity was defined in two versions *TCI* and *TC* (abbreviation for *topological complexity*). The vertex degrees in *TC* are taken from the molecular graph  $G$ , whereas in *TCI* they are taken from the corresponding subgraphs  $G_i$ . All overall indices  $OI(G)$  are also presented in vector form  $OI'(G)$  incorporating all  $e$ th order terms  ${}^eOI(G)$  where  $I = C, W, M1$  or  $M2$  indicates the types of the overall index. The current value  $e$  of the number of edges in the subgraph runs from zero for zero-order complexity (complexity of vertices), to one for first-order complexity (complexity of edges), etc., to  $e = E$  for the entire graph having  $E$  edges.

$$TC'(G) = TC\{{}^0TC, {}^1TC, {}^2TC, \dots, {}^E TC\}; \quad TC(G) = \sum_{e=0}^E {}^eTC(G) = \sum_{i=1}^K A_i(G_i \subset G) \quad (4.23)$$

$$OW'(G) = OW\{{}^0OW, {}^1OW, {}^2OW, \dots, {}^EOW\}; \quad OW(G) = \sum_{e=1}^E {}^eOW(G) = \sum_{i=1}^K W_i(G_i \subset G) \quad (4.24)$$

$$OM'(G) = OM\{{}^0OM, {}^1OM, {}^2OM, \dots, {}^EOM\}; \quad OM(G) = \sum_{e=1}^E {}^eOM(G) = \sum_{i=1}^K M_i(G_i \subset G) \quad (4.25)$$

In equations (4.23) and (4.25) we omit for brevity additional symbols used to distinguish between the two overall connectivities  $TC$  and  $TC1$ , and between the first and second Zagreb index,  $M1$  and  $M2$ , respectively.

All of the three types of overall topological indices satisfy the requirement for a complexity measure. They increase with both size and connectedness of the structure, and are very sensitive toward subtle complexity patterns in acyclic and cyclic structures. They increase with the number and size of the branches and cycles, as well as with their more central location (a topological feature called *centrality*), and with the closer branch/branch or cycle/cycle location (features called *branch adjacency* and *cycle adjacency*, respectively). Some of these complexity patterns are illustrated in Figures 4.4 and 4.5 (vide infra), along with the information indices on substructure distributions, which we will define below.

The partitioning of the overall topological indices into their substructural components (equations 4.23–4.25) gives rise to a typical situation for applying our magnitude-based information-theoretic approach. Each of the three types of overall indices may thus be supplemented by an *overall information index*, defined on the set of  $e$ th-order terms. Due to the detailed description of the topological structure provided by these parent sets, one might anticipate these to be the best structural complexity indices that a Shannon-type function could produce. The *overall connectivity information indices*,  ${}^mI(TC)$  and  ${}^mI(TC1)$ , the *overall Wiener information index*,  ${}^mI(OW)$ , and the *overall Zagreb information indices*,  ${}^mI(OM1)$  and  ${}^mI(OM2)$  are defined below along with the *substructure count information index*,  ${}^mI(K)$ :

$${}^mI(K) = K \log_2 K - \sum_{e=0}^E e K \log_2 {}^eK; \quad {}^mI_{av}(K) = {}^mI(K) / K \quad (4.26)$$

$${}^mI(TC) = TC \log_2 TC - \sum_{e=0}^E e TC \log_2 {}^eTC; \quad {}^mI_{av}(TC) = {}^mI(TC) / TC \quad (4.27)$$

$${}^mI(TC1) = TC1 \log_2 TC1 - \sum_{e=1}^E e TC1 \log_2 {}^eTC1; \quad {}^mI_{av}(TC1) = {}^mI(TC1) / TC1 \quad (4.28)$$

The summation in equation (4.28) does not start from  $e = 0$ , because the vertex degrees in  $TC1$  are those in the corresponding subgraphs, and for isolated vertices  $a_i = 0$ . Similarly,  ${}^0OW = 0$  in equation (4.29) below because the distances between the isolated vertices are zero. The vertex degrees in the overall Zagreb information indices (equations 4.30 and 4.31) are those from the entire graph, and the second overall Zagreb index has a zero-

Table 4.3: Overall topological and overall information indices of the graphs from Figure 4.6

Indices	Graphs				
	3	4	5	6	19
$K$	15	17	20	54	57
${}^mI(K)$	32.24	43.32	64.16	138.20	144.72
$TC1$	40	64	110	310	326
${}^mI(TC1)$	78.84	115.92	232.99	686.34	721.67
$TC$	60	100	160	482	522
${}^mI(TC)$	135.81	209.22	232.20	1136.39	1225.71
$OW$	56	80	190	510	483
${}^mI(OW)$	101.32	137.60	335.04	994.51	975.92
$OM1$	110	292	320	1228	1407
${}^mI(OM1)$	247.71	600.23	752.71	2897.50	3301.05
$OM2$	60	68	342	2015	2025
${}^mI(OM2)$	135.39	143.91	709.11	4131.10	4409.24

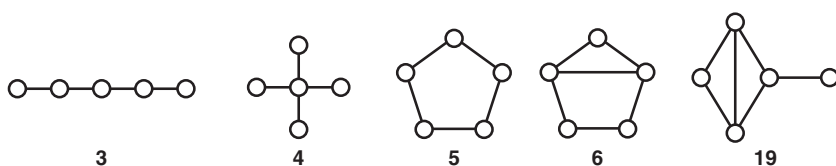


Figure 4.6: Five-vertex acyclic and cyclic graphs ordered according to their increasing complexity.

order term introduced by analogy with the zero-order term of the overall connectivity index  $TC$  as the sum of all vertex degrees in the graph.

$${}^mI(OW) = OW \log_2 OW - \sum_{e=1}^E {}^eOW \log_2 {}^eOW; \quad {}^mI_{av}(OW) = {}^mI(OW)/OW \quad (4.29)$$

$${}^mI(OM1) = OM1 \log_2 OM1 - \sum_{e=0}^E {}^eOM1 \log_2 {}^eOM1; \quad {}^mI_{av}(OM1) = {}^mI(OM1)/OM1 \quad (4.30)$$

$${}^mI(OM2) = OM2 \log_2 OM2 - \sum_{e=0}^E {}^eOM2 \log_2 {}^eOM2; \quad {}^mI_{av}(OM2) = {}^mI(OM2)/OM2 \quad (4.31)$$

Table 4.3 summarizes the values of the overall topological indices  $K$ ,  $TC$ ,  $TC1$ ,  $OW$ ,  $OM1$ , and  $OM2$  for graphs **3–6** and **19** from Figure 4.6. This is a modified Figure 4.3 with the totally disconnected graph **2** no longer shown because of its zero values for all indices. The complete graph **7** as the most complex graph having five vertices is characterized with very high values of all substructure-based indices, and is also not shown in Figure 4.6. Instead, the complexity of a new five-vertex graph **19** is analyzed with the expectation that it will be more complex than the other bicyclic graph **6**, because it has a branch as an additional complexity factor.

The expected ordering of these five graphs with their complexity increasing from the linear graph **3** to the star graph **4** to the monocyclic graph **5** to the bicyclic graph **6** to the bicyclic branched graph **19** is:

$$\mathbf{3} \Rightarrow \mathbf{4} \Rightarrow \mathbf{5} \Rightarrow \mathbf{6} \Rightarrow \mathbf{19}$$

As seen from Table 4.3, this ordering is confirmed by all indices examined. The only exception is found for the overall Wiener index and its information counterpart, which reorder graphs **6** and **19**. The bicyclic branched graph **19** has three subgraphs more than the unbranched bicyclic graph **6**. However, these are small size subgraphs having two and three edges per subgraph, and the increase in the corresponding  $OW$  terms they provide does not suffice to compensate for the larger number of longer linear subgraphs with four and five edges in graph **6**. Indeed, there is a competition between the two complexity factors of branching and cycle size in these two graphs. The overall Wiener index and the overall Wiener information indices appear to be the only ones to favor branching over cycle size.

A more detailed analysis of acyclic branching is made by comparing the complexity of the five isomeric acyclic hexanes (Figure 4.4 and Table 4.4). The magnitude-based overall information indices, as well as their parent overall topological indices, reflect the regular increase in complexity in the sequence

$$\mathbf{8} \Rightarrow \mathbf{9} \Rightarrow \mathbf{10} \Rightarrow \mathbf{11} \Rightarrow \mathbf{12}$$

Several complexity factors are thus correctly reflected: the increase in the number of branches (**8**  $\Rightarrow$  **9,10**  $\Rightarrow$  **11,12**), the branching at a vertex of higher degree (**11**  $\Rightarrow$  **12**), and the shifting of a branch toward a more central position (**9**  $\Rightarrow$  **10**). Notably, the change in the overall information indices upon the action of centrality factor (the **9**  $\Rightarrow$  **10** transformation) is

Table 4.4: Overall topological and overall information indices of the five acyclic graphs **8–12** (Figure 4.4)

Indices	Graphs				
	8	9	10	11	12
$K$	21	24	25	28	30
${}^mI(K)$	50.36	58.56	61.0	67.48	72.78
$TC1$	70	88	94	112	122
${}^mI(TC1)$	159.55	195.47	206.01	239.83	258.03
$TC$	100	127	136	164	181
${}^mI(TC)$	251.61	311.84	329.63	388.48	422.82
$OW$	126	154	161	188	197
${}^mI(OW)$	263.92	315.14	328.32	373.74	392.74
$OM1$	188	277	300	404	505
${}^mI(OM1)$	471.24	675.13	720.40	947.92	1165.91
$OM2$	130	149	161	172	168
${}^mI(OM2)$	321.65	358.18	383.02	399.97	389.94

larger than that in the corresponding parent overall topological indices. This illustrates the high sensitivity of the newly introduced information indices toward subtle complexity patterns.

The ordering of the five hexanes coincides with the ordering produced by other symmetry-independent indices, whereas not a single symmetry-based index (including their best representative, the Bertz index [24]) is able to mirror all of the three branching patterns identified in this series of graphs [38]. The only partial reordering that is produced by the twelve overall topological and information indices deals with graphs **11** and **12**. The second overall Zagreb index  $OM2$  and its information analogue  ${}^mI(OM2)$  fail to show graph **12** as more complex than graph **11**. The reason for this result can be traced to the smaller index increment produced by the  $OM2$  function for two neighboring vertices of degree four and two, as compared to two neighbors of degree three ( $3 \times 3 > 4 \times 2$ ). Thus, the  $OM2$  and  ${}^mI(OM2)$  indices cannot always reflect the greater contribution to molecular complexity coming from branching at an atom of higher valence.

## 4.7 Concluding Remarks

This study revises the earlier criticism of the present author [11–13, 55] toward the use of the Shannon information function as a complexity

measure. The findings in our previous studies were that the information content of a molecule, calculated by the standard symmetry- or equivalence-based scheme, is a good measure for molecular diversity, particularly for the atomic composition of molecules. At the same time, it was concluded that no information-theoretic index constructed on such a basis could measure the structural complexity of a molecule or any other system. The only complexity index of this type with relatively good performance is the Bertz index [24] based on graph connections. However, as shown in this chapter, the parent symmetry-based information index on the graph connections is a very poor complexity measure. The success of the Bertz index is due to the very essential addition of a "size term," which in most cases compensates for the weaknesses inherent for the information-theoretic technique employed. The correction incorporated is not based on a rigorous theory and has only practical importance. Yet, even this index faces difficulties in reflecting some subtle complexity patterns in acyclic and cyclic systems.

The second information-theoretic scheme, developed by Bonchev and Trinajstić [25] in 1977, is based on the partitioning of certain magnitudes or weights of the system into the contributions of its elements, and is thus not related to the symmetry of the system. The magnitude-based information indices on graph distances,  ${}^mI(\text{distance})$ , introduced in that study mirror to a high extent molecular complexity patterns in *isomeric* acyclic and cyclic systems. When combined with another index to account for the size of the system, this index has found application in structure-property and structure-activity relationships (QSPR, QSAR) studies [66, 67]. However, the  ${}^mI(\text{distance})$  index does not yield a good complexity measure for systems of different size, owing to the opposing trends of increasing with size and decreasing with branching and cyclicity. For this reason, the conclusion was drawn in our previous work [11–13, 55] that the Shannon information function could not be a measure of structural complexity.

In the present study, these rumors of the premature death of information-theoretic complexity measures were found to be strongly exaggerated, though they have gained further ground specifically for the *symmetry-based* information indices. The revision refers to the *magnitude-based* information complexity measures, provided the graph invariant used to construct the magnitude distribution is selected so as *to increase* with both size and structural complexity factors. We have proposed in this chapter several such indices, beginning with the  ${}^mI(\text{conn deg})$  index. It is based on the distribution of graph vertices according to the number of connections (two-edge subgraphs) that emanate from each vertex. Indeed, the number

of graph connections is not a universal graph invariant, and cannot serve as a basis for a universal information-theoretic complexity measure. It cannot, for example, distinguish graphs in which all vertices belong to a single orbit. This indicated that the distributions of more complex subgraphs having three or more edges should be addressed as well. Instead of these potentially useful but only partial solutions to the problem, we turned to the complete structural characterization of a graph by the magnitude-type distribution of all subgraphs in the graph, as well as by a similar distribution provided by the recently developed overall topological indices [58, 61–63]. The latter ascribe certain weights or magnitudes to each of the subgraphs, such as the subgraph total adjacency, [58, 61] its total distance, [62] and the vertex-degree-based Zagreb indices [63]. The overall information indices thus constructed and, particularly, the two information overall connectivity indices and that on the total substructure count would seem to satisfy to a high degree the requirements for a good measure of structural complexity.

#### 4.8 References

1. O.N. Temkin, A.V. Zeigarnik, and D. Bonchev, *Chemical Reaction Networks. A Graph Theoretical Approach*, CRC Press, Boca Raton, FL, 1996.
2. O.N. Temkin, A.V. Zeigarnik, and D. Bonchev, Graph-Theoretical Models of Complex Reaction Mechanisms and Their Elementary Steps, in: *Graph-Theoretical Approaches to Chemical Reactivity*, D. Bonchev and O. Mekenyan (Eds), Kluwer Academic, Dordrecht, 1994, pp. 143–278.
3. K. Gordeeva, D. Bonchev, D. Kamenski, and O.N. Temkin, Enumeration, Coding, and Complexity of Linear Reaction Mechanisms, *J. Chem. Inf. Comput. Sci.*, **34**, 244–247 (1994).
4. C. Shannon and W. Weaver, *Mathematical Theory of Communication*, University of Illinois Press, Urbana, MI, 1949.
5. S.M. Dancoff, and H. Quastler, The Information Content and Error Rate of Living Things, in: *Essays on the Use of Information Theory in Biology*, H. Quastler (Ed.), University of Illinois Press, Urbana, MI, 1953.
6. H. Linshitz. The Information Content of a Bacterial Cell. In: *Essays on the Use of Information Theory in Biology*, H. Quastler (Ed.), University of Illinois Press, Urbana, MI, 1953.

7. N. Rashevsky, Life, Information Theory, and Topology. *Bull. Math. Biophys.*, **17**, 229–235 (1955).
8. L. Brillouin, *Science and Information Theory*, Academic Press, New York, 1956.
9. A. Mowshowitz, Entropy and the Complexity of Graphs: I. An Index of the Relative Complexity of a Graph, *Bull. Math. Biophys.*, **30**, 175–204 (1968).
10. D. Bonchev, Information Indices for Atoms and Molecules. *MATCH*, **7**, 65–113 (1979).
11. D. Bonchev and D.E. Polansky, On the Topological Complexity of Chemical Systems, in: *Graph Theory and Topology in Chemistry*, R.B. King and D.H. Rouvray (Eds), Elsevier, Amsterdam, 1987, pp. 126–158.
12. D. Bonchev, The Problems of Computing Molecular Complexity, in: *Computational Chemical Graph Theory*, D.H. Rouvray (Ed.), Nova Publications, New York, 1990, pp. 34–67.
13. D. Bonchev and W.A. Seitz, The Concept of Complexity in Chemistry, in: *Concepts in Chemistry: A Contemporary Challenge*, D.H. Rouvray (Ed.), Research Studies Press, Taunton, UK, 1996, pp. 348–376.
14. D. Bonchev, Information Theory Interpretation of the Pauli Principle and Hund Rule, *Intern. J. Quantum Chem.*, **19**, 673–679 (1981).
15. B. Rousseva and D. Bonchev, A Theoretic-Information Variant of Nuclide Systematics. *Commun. Math. Chem (MATCH)*, **4**, 173–192 (1978).
16. D. Bonchev, V. Kamenska, D. Kamenski, Informationsgehalt Chemischer Elemente, *Monatsh. Chem.*, **108**, 477–487 (1977).
17. D. Bonchev and V. Kamenska, Information Theory in Describing the Electronic Structure of Atoms, *Croat. Chem. Acta*, **51**, 19–27 (1978).
18. D. Bonchev and V. Kamenska, Predicting the Properties of the 113–120 Transactinide Elements, *J. Phys. Chem.*, **85**, 1177–1186 (1981).
19. D. Bonchev, D. Kamenski, and V. Kamenska, Symmetry and Information Content of Chemical Structures, *Bull. Math. Biol.*, **38**, 119–133 (1976).
20. E. Trucco, A Note on the Information Content of Graphs, *Bull. Math. Biophys.*, **18**, 129–135 (1956); E. Trucco, On The Information Content of Graphs: Compound Symbols; Different States for Each Point, *Bull. Math. Biophys.*, **18**, 237–253 (1956).
21. D. Bonchev, O. Mekenyan, and N. Trinajstić, Isomer Discrimination by Topological Information Approach, *J. Comput. Chem.*, **2**, 127–148 (1981).

22. M. Gordon and J.W. Kennedy, *J. Chem. Soc. Faraday Trans. II*, **69**, 484–504 (1973).
23. J. R. Platt, Prediction of Isomeric Differences in Paraffin Properties, *J. Phys. Chem.*, **56**, 328–336 (1952).
24. S. Bertz, The First General Index of Molecular Complexity, *J. Am. Chem. Soc.*, **103**, 3599–3601 (1981).
25. D. Bonchev, and N. Trinajstić, Information Theory, Distance Matrix, and Molecular Branching, *J. Chem. Phys.*, **67**, 4517–4533 (1977).
26. D. Bonchev and N. Trinajstić, Chemical Information Theory. Structural Aspects, *Intern. J. Quantum Chem. Symp.*, **16**, 463–480 (1982).
27. D. Bonchev, A. T. Balaban and O. Mekenyan, Generalization of the Graph Center Concept, and Derived Topological Indexes. *J. Chem. Inf. Comput. Sci.*, **20**, 106–113 (1980).
28. D. Bonchev, A.T. Balaban and M. Randić, The Graph Center Concept for Polycyclic Graphs, *Intern. J. Quantum Chem.*, **19**, 61–82 (1981).
29. F. Harary, *Graph Theory*. Addison-Wesley, Reading, MA, 1969.
30. A.T. Balaban, Highly Discriminating Distance-Based Topological Index, *Chem. Phys. Lett.*, **89**, 399–404 (1982).
31. D. Bonchev, O. Mekenyan and A.T. Balaban, An Iterative Procedure for the Generalized Graph Center in Polycyclic Graphs, *J. Chem. Inf. Comput. Sci.*, **29**, 91–97 (1989).
32. S.C. Basak, A.B. Roy and J.J. Ghosh, Study of the Structure-Function Relationship of Pharmacological and Toxicological Agents Using Information Theory, in: *Proceedings of the Iind International Conference on Mathematical Modeling*, X, J.R. Avula *et al.* (Eds), University of Missouri, Rolla, Missouri, Vol. II, pp. 851–856.
33. V.R. Magnuson, D.K. Harris and S.C. Basak, Topological Indices Based on Neighborhood Symmetry: Chemical and Biological Applications, in: *Chemical Applications of Topology and Graph Theory*, R.B. King (Ed.), Elsevier, The Netherlands, pp. 178–191.
34. S.C. Basak, Information Theoretic Indices of Neighborhood Complexity and Their Applications, in: *Topological Indices and Related Descriptors in QSAR and QSPR*, J. Devillers and A.T. Balaban (Eds), Gordon and Breach, Reading, UK, pp. 563–593.
35. D. Bonchev, *Information-Theoretic Indices for Characterization of Chemical Structures*. Research Studies Press, Chichester, UK, 1983.
36. S.H. Bertz, Complexity of Molecules and Their Synthesis, in: *Mathematical Chemistry Series, Vol. VII Complexity in Chemistry*, D.

- Bonchev and D.H. Rouvray (Eds), Taylor & Francis, London, UK, Chapter 3, herein.
37. S.H. Bertz and C.M. Zamfirescu, New Complexity Indices Based on Edge Covers, *Commun. Math. Chem. (MATCH)*, **42**, 9–70 (2000).
  38. S. Nikolić, N. Trinajstić, I.M. Tolić, G. Rücker, and Ch. Rücker, On Molecular Complexity Indices, in: *Mathematical Chemistry Series, Vol. VII Complexity in Chemistry*. D. Bonchev and D.H. Rouvray (Eds), Taylor & Francis, London, UK, Chapter 2, herein.
  39. H. Wiener, Structural Determination of Paraffin Boiling Points, *J. Am. Chem. Soc.*, **69**, 17–20 (1947).
  40. D. Bonchev, O. Mekenyan, and N. Trinajstić, Topological Characterization of Cyclic Structures, *Intern. J. Quantum Chem.*, **17**, 845–893 (1980).
  41. A.T. Balaban, D. Bonchev, X. Liu, and D.J. Klein, Molecular Cyclicity and Centricity of Polycyclic Graphs. I. Cyclicity Based on Resistance Distances or Reciprocal Distances, *Int. J. Quantum Chem.*, **50**, 1–20 (1994).
  42. E. Ruch and I. Gutman, The Branching Extent of Graphs, *J. Combinatorics*, **4**, 285–295 (1979).
  43. S.H. Bertz, Branching in Graphs and Molecules, *Discrete Appl. Math.*, **19**, 41–70 (1988).
  44. D. Bonchev, Topological Order in Molecules. 1. Molecular Branching Revisited, *Theochem*, **336**, 137–156 (1995).
  45. Randić, M. On Molecular Branching, *Acta Chim. Slovenica*, **44**, 57–77 (1997).
  46. D. Bonchev, Shannon's Information in Assessing Structural Complexity, to be published.
  47. D. Minoli, Combinatorial Graph Complexity, *Atti Acad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.* (Ser. 8), **59**, 651–661 (1975).
  48. D. Bonchev, O.N. Temkin, and D. Kamenski, On the Complexity of Linear Reaction Mechanisms, *React. Kinet. Catal. Lett.*, **15**, (1980) 119–124.
  49. D. Bonchev, D. Kamensky, and O.N. Temkin, Complexity Index for the Linear Mechanisms of Chemical Reactions, *J. Math. Chem.*, **1**, 345–388 (1987).
  50. I. Gutman, R.B. Mallion, and J.W. Essam, Counting the Spanning Trees of a Labeled Molecular Graph, *Mol. Phys.*, **50**, 859–877 (1983).
  51. P.E. John, R.B. Mallion, and I. Gutman, An Algorithm for Counting Spanning Trees in Labeled Molecular Graphs Homeomorphic to Cata-condensed Systems, *J. Chem. Inf. Comput. Sci.*, **38**, 108–112 (1998).

52. S.H. Bertz, W.C. Herndon, Similarity of Graphs and Molecules, in: *Artificial Intelligence Applications in Chemistry*, American Chemical Society, Washington, DC, 1986, pp. 169–175.
53. S.H. Bertz and T.J. Sommer, Rigorous Mathematical Approaches to Strategic Bonds and Synthetic Analysis Based on Conceptually Simple New Complexity Indices, *Chem. Commun.*, 2409–2410 (1997).
54. S.H. Bertz and W.F. Wright, The Graph Theory Approach to Synthetic Analysis: Definition and Application of Molecular Complexity and Synthetic Complexity, *Graph Theory Notes of New York* (NY Acad. Sci.), **XXXV**, 32–48 (1998).
55. D. Bonchev, Kolmogorov's Information, Shannon's entropy, and Topological Complexity of Molecules, *Bulg. Chem. Commun.*, **28**, 567–582 (1995).
56. D. Bonchev, Novel Indices for the Topological Complexity of Molecules. *SAR QSAR Environ. Res.*, **7**, 23–44 (1997).
57. D. Bonchev, Overall Connectivity and Topological Complexity: A New Tool for QSPR/QSAR, in: *Topological Indices and Related Descriptors in QSAR and QSPR*, J. Devillers and A.T. Balaban (Eds), Gordon and Breach, The Netherlands 1999, pp. 361–401.
58. D. Bonchev, Overall Connectivities/Topological Complexities: A New Powerful Tool for QSPR/QSAR, *J. Chem. Comput. Sci.*, **40**, 934–941 (2000).
59. R.G.A. Bone and H.O. Villar, Exhaustive Enumeration of Molecular Substructures, *J. Comput. Chem.*, **18**, 86–107 (1997).
60. G. Rücker and Ch. Rücker, Walk Counts, Labyrinthicity and Complexity of Acyclic and Cyclic Graphs and Molecules, *J. Chem. Inf. Comput. Sci.*, **40**, 99–106 (2000).
61. D. Bonchev, Overall Connectivity—A Next Generation Molecular Connectivity, *J. Mol. Graphics Model.*, **5271**, 1–11 (2001).
62. D. Bonchev, The Overall Wiener Index—A New Tool for Characterization of Molecular Topology, *J. Chem. Inf. Comput. Sci.*, **41**, 582–592 (2001).
63. D. Bonchev, N. Trinajstić, Overall Molecular Descriptors. 3. Overall Zagreb Indices, *SAR QSAR Environ. Res.*, **12**, 213–235 (2001).
64. I. Gutman, B. Rušćić, N. Trinajstić and C.W. Wilcox, Jr., Graph Theory and Molecular Orbitals. 12. Acyclic Polyenes, *J. Chem. Phys.*, **62**, 3399–3405 (1975).
65. S. Nikolić, I.M. Tolić, N. Trinajstić and I. Baučić, On the Zagreb Indices As Complexity Indices, *Croat. Chem. Acta*, in press.
66. O. Mekenyan, D. Bonchev and N. Trinajstić, Chemical Graph Theory Modeling the Thermodynamic Properties of Molecules, *Intern. J.*

- Quantum Chem. Symp.*, **18**, 369–380 (1980).
67. D. Bonchev, W.A. Seitz, C.F. Mountain and A.T. Balaban, Modeling the Anticarcinogenic Action of Some Retinoid Compounds by Making Use of the OASIS Method. III. Inhibition of the Induction of Ornithine Decarboxylase by Arotinoids, *J. Med. Chem.*, **37**, 2300–2307 (1994).