From Molecular to Biological Structure and Back†

Danail Bonchev* and Gregory A. Buck

Center for the Study of Biological Chemistry, Virginia Commonwealth University, P. O. Box 842030, Richmond, Virginia 23284-2030

Received February 15, 2007

A comparative analysis of the topological structure of molecules and molecular biology networks revealed both similarity and differences in the methods used, as well as in the essential features of the two types of systems. Molecular graphs are static and, due to the limitations in atomic valence, show neither power distribution of vertex degrees nor "small-world" properties, which are typical for dynamic evolutionary networks. Areas of mutual benefits from an exchange of methods and ideas are outlined for the two fields. More specifically, chemical graph theory might make use of some new descriptors of network structure. Of interest for quantitative structure−property relationship/quantitative structure−activity relationship and drug design might be the conclusion that descriptors based on distributions of vertex degrees, distances, and subgraphs seem to be more relevant to biological information than the single-number descriptors. The network concepts of centrality, clustering, and cliques provide a basis for similar studies in theoretical chemistry. The need of dynamic theory of molecular topology is advocated.

1. INTRODUCTION

Theoretical chemistry has benefited a lot during the past 40 years from applying the methods of graph theory[1,2] and information theory[3,4] to characterize molecular structure. The quantitative relationships between molecular descriptors and the properties and biological activity of chemical compounds (quantitative structure−property and structure−activity relationships; QSPR/QSAR)[5] have found applications in areas such as drug design, development of new chemical products, polymer chemistry, chemical documentation, and so forth.[6] The progress in chemical graph theory and chemical information theory has been achieved to a high extent independently from other applied areas of these two theories.

Recently, networks became of great interest as a common language for characterizing complex dynamic systems in nature, technology, and society in a holistic manner, considering the system as a whole. These developments were most pronounced in biology, which in its postgenomic era is focused on the biochemical networks formed in the cell by genes, proteins, and metabolites. The explosively expanding fields of systems biology, bioinformatics, and computational biology have attracted scientists and expertise from graph theory and information theory applications in such areas as physics, chemistry, computer sciences, and social sciences. The considerable progress achieved in characterizing and modeling networks in these specific branches of science contributed greatly to the revolution in biological science by creating a wealth of network-based concepts, methods, and software for elucidating the mechanisms of the living cell.[7−15]

This article is aiming to provide a comparative analysis of molecular structure and biological structure at the level of intracellular biochemical networks, outlining areas of similarity and dissimilarity, as well as areas of mutual benefit from the exchange of concepts and methods.

2. MOLECULAR GRAPHS VERSUS BIOLOGICAL INTRACELLULAR NETWORKS

Table 1 summarizes the types of graphs and their elements used in describing the topology of molecules and the variety of molecular biology networks, some basic terms used, and their specific meaning in both areas.

Some of the terms used in Table 1 need additional elucidation. Integrated are those intracellular networks in which the links stand for more than one type of interaction. A typical example is transcriptional regulatory network, the link in which may mean either protein−protein interaction or protein−gene regulatory interaction, and in some more detailed versions, even post-translational modification (phosphorylation) is included. Integrated intracellular networks may incorporate all possible types of interactions within the cell. Only few such networks are published so far, due to their very complex nature. However, this is the next frontier in the systems biology approach to elucidation of the organization and functioning of the living cell. Software assisting the construction and analysis of such networks is already available.[16]

Networks in which the nodes represent a group of ingredients could be of structural or functional type. On a structural level, groups may be modules (clusters), the nodes in which are well-connected between themselves but have less links with the remaining network nodes.[17−19] On a functional level, these are groups of ingredients jointly performing a certain biological action. Metabolic pathways are such natural groupings of metabolites, and a network focused on the links between all pathways was recently proposed. Other examples are the groups of proteins performing joint action. Such higher hierarchical-level networks

† Dedicated to Professor Nenad Trinajstić on the occasion of his 70th birthday.

* Corresponding author phone: (804) 827-7375; fax: 804-828-1961; e-mail: dgbonchev@vcu.edu.

10.1021/ci7000617 CCC: $37.00
Published on Web 04/04/2007

Table 1. Comparison of Molecular Graphs and Biological Networks, and Specific Meaning of Terms in Both Fields

molecular graphs	biological networks
term/object described	term/object described
graph/molecule	network/metabolic, transcriptional regulatory, signaling, protein−protein interaction, pathway−pathway interaction, integrated, etc.
subgraph/molecular fragment	subnetwork/pathway, structural and functional modules and motifs
vertex/atom (atomic nucleus)	node/gene, protein, metabolite, small molecule, and groups of those
edge/chemical bond	link/binding, regulation, coexpresion, genetic, posttranslational modification, molecular transport, etc.; sharing (for groups of genes, proteins, or metabolites)
undirected graph/molecule	undirected network/protein−protein interaction networks; networks with sharing of genes, proteins, or metabolites
	directed networks/transcriptional regulatory, signaling, integrated networks
connected graphs used only/common molecular graphs	connected and disconnected networks/some proteins, genes, and metabolites may remain idle
nonweighted graphs/common molecular graphs	nonweighted networks/common biological networks
weighted graphs/charges, electronegativities, bond orders, etc.	weighted networks/expression data, number of shared ingredients
homogeneous graphs (based on single interaction type)/chemical bonding	both homogeneous networks/protein−protein interaction and metabolic networks and heterogeneous/gene regulatory, signaling, integrated networks
static graph/unchanged molecular structure and composition	dynamic evolving network/addition and deletion of nodes and links

are typically sharing ingredients (metabolites or proteins) that can perform more than one biological action. Surprisingly, it was found that at least 50% of proteins in yeast can perform two or more functions and that there are no such biological functions that can be performed only by highly specialized groups of proteins.[20]

The biochemical networks show much larger variety of graph types than molecular structure. Thus, molecular graphs are undirected, whereas in some types of biological networks, interactions are directed (metabolic, regulation, signaling, etc.). One might consider a version of molecular graphs with directed edges between atoms of different electronegativity. Such graphs would necessarily be weighted as well, accounting for the magnitude of electronegativity. Molecular graphs are homogeneous; the edges in them stand for a single type of interaction. Some biological networks are heterogeneous and include more than one type of node (e.g., genes and proteins) and more than one type of interaction (e.g., binding and regulation).

The most significant difference between biological networks and molecular graphs is that the biological network structure is dynamic; it evolves or degrades by adding or deleting nodes and links, as a result of mutations. Molecular structure as described by molecular graphs is static, and this is a serious limitation for predicting molecular behavior. The only exception is the area of polymer studies, in which the accounting for dynamics resulted in deriving equations relating the Wiener number of the polymer to its radius of gyration and viscosity.[21,22] Perhaps, new types of molecular graphs are needed for revealing the dynamics of chemical interactions. The graphs, recently defined from the valence electrons of atoms,[23] as well as the assignment of wave functions to graphs[24] may be considered as an important step in the right direction.

3. CONNECTIVITY IN MOLECULAR GRAPHS AND BIOLOGICAL NETWORKS

The basic connectivity descriptors used in both fields are almost identical, indeed. For a graph/network having V vertices/nodes and E edges/links, the connectivity descriptors are defined from the adjacency matrix $\mathbf{A}$, the elements a_{ij} of which are equal to 1 for i and j − adjacent, and equal to 0 otherwise. The local connectivity is described by the vertex degree a_i, and the global connectivity is characterized either by the total adjacency A or by the average vertex degree $<a_i>$. While total adjacency is rarely used for biological networks, due to their huge size, for network comparison purposes, the average vertex degree is used in parallel with *network connectedness*[25] Conn(G), called also *network density d*(G). This is the total adjacency doubly normalized, so as to be defined within the 0 to 1 range:

$$a_i = \sum_{i=1}^{V} a_{ij} \tag{1a}$$

$$A = \sum_{i=1}^{V} a_i = \sum_{i=1}^{V} \sum_{j=1}^{V} a_{ij} \tag{1b}$$

$$<a_i> = \frac{A}{V} \tag{1c}$$

$$\mathrm{Conn}(G) = \frac{A}{V(V-1)} \tag{2a}$$

or

$$\mathrm{Conn}'(G) = \frac{A}{V^2} \tag{2b}$$

Equation 2b is used in cases of networks having nonzero entries in the adjacency matrix main diagonal, originating from self-interaction (e.g., protein dimer formation).

Connectedness has not been used before in characterizing adjacency in molecular graphs of different size. Vice versa, another descriptor introduced to characterize molecular topology started being applied as a measure of network complexity.[26,27] This is the information theoretic index for vertex degree distribution, which makes use of Shannon's

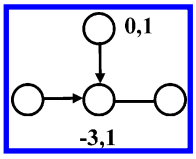

Figure 1. Illustration to the definition of in-degrees (the first digit) and out-degrees (the second digit) of the nodes of the directed network.

equations for average and total information content $<I_{vc}>$ and I_{vc}, respectively:

$$<I_{vc}> = -\sum_{i=1}^{V} p_i \log_2 p_i \tag{3a}$$

$$I_{vc} = A \log_2 A - \sum_{i=1}^{V} a_i \log_2 a_i \tag{3b}$$

where the logarithm at base 2 is taken to measure the information in bits (eq 3b), and in bits/vertex (eq 3a), and $p_i = a_i/A$ is the probability of a randomly chosen atom i to have degree a_i.

The above analysis refers to molecular graphs and undirected networks. In directed networks, separate incoming (or shortly "in-") and outgoing (or shortly "out-") degrees are defined depending on the direction of network links. More specifically, the binary relation between two vertices becomes −1 for incoming edges and +1 for the outgoing ones (see Figure 1). Thus, eqs 1−3 are calculated separately for the in- and out-degrees.

While there is not a big difference in the descriptors used for molecular graphs and biological networks, there is an essential difference in the ways vertex degrees are distributed in them. The degree distribution in molecular graphs has a very low upper limit, determined by the maximal atomic valence. As a result, the most frequent degree values are within the range of 1−4. In a series of isomeric compounds, the vertex degree distribution follows the Poisson distribution, typical for random networks (Figure 2a).

In contrast, there is no such limitation in molecular biology networks, where the highest node degree may exceed 100. Such highly connected nodes were termed "hubs". They are viewed as essential for cell survival. The deletion of such a highly connected protein or gene is usually lethal for the cell. The node degree distribution in biological networks is not random. As shown in the seminal paper of Barabaśi and Albert,[11] this distribution follows the power law:

$$p(k) = bk^{-\gamma} \tag{4}$$

Here, b is a numerical coefficient, and the exponent γ for biological networks has values within the 2.0−3.0 range (Figure 2b). As a consequence of the power law distribution,

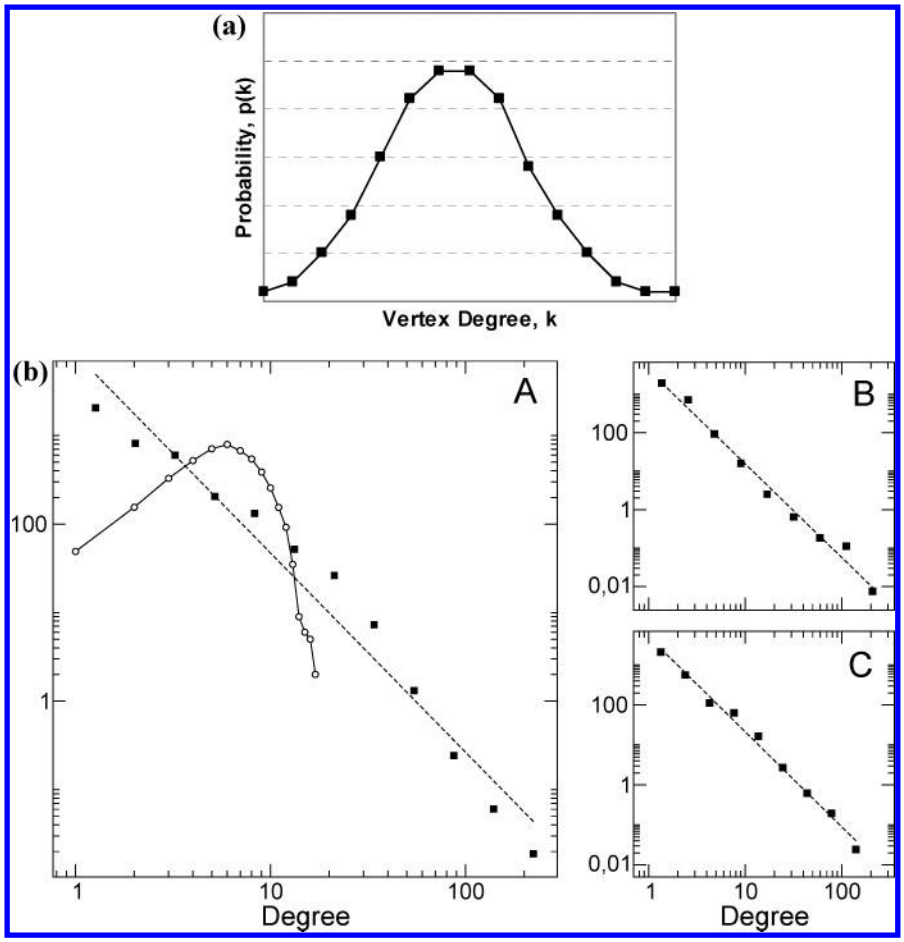

Figure 2. (a) Poisson distribution of the vertex degree in a series of molecular graphs of the same size. (b) A. Example of power law distribution, which is linear in log/log coordinates.[28] The Poisson distribution for the same system is also shown as a curve with a maximum. B and C. The power law distribution for the degrees in the protein−protein interaction networks of (B) *C. elegans* ($V = 3280$; $E = 4549$; $\gamma = 2.43$) and (C) *S. cerevisiae* ($V = 3228$; $E = 5625$; $\gamma = 2.37$).

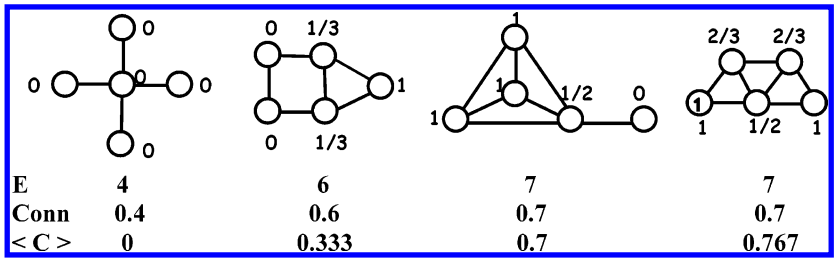

Figure 3. Four graphs with five vertices, their total number of edges *E*, connectedness Conn (calculated by eq 2a), average clustering coefficient $<C>$ (calculated by eq 5b), and vertex clustering coefficients shown at each vertex (calculated by eq 5a). As demonstrated, clustering around a vertex is possible only in tri-membered cycles. In all other structures, there are no edges between the vertex first neighbors (eq 5a; $E_i = 0$).

biological networks contain few highly connected nodes, and many nodes of degree 1 or 2.

4. CLUSTERING COEFFICIENT

Vertex degree is not the only measure of connectivity on a local level of graphs and networks. It is important also to know whether and to what extent the first neighbors of a node are linked between themselves. In the case of sufficient first-neighbors interconnectivity, the graph has a clustered structure, which is quantified by the clustering coefficient, C_i.[28] This coefficient normalizes, within the range 0−1, the number of edges E_i between the first neighbors of the vertex i by dividing it by the maximal number of such edges, $a_i(a_i - 1)/2$. The overall degree of network clustering is assessed by averaging the clustering coefficients of all vertices:

$$C_i = \frac{2E_i}{a_i(a_i - 1)} \qquad 0 \le C_i \le 1 \tag{5a}$$

$$<C> = \frac{\sum_{i=1}^{V} C_i}{V} \tag{5b}$$

Equations 5a and 5b are exemplified in Figure 3 with four five-vertex graphs ordered according to their connectedness Conn, which increases with the number of edges and cycles. It is seen that the clustering coefficients of all atoms in acyclic graphs are equal to zero, and the same is valid for vertices in acyclic branches of cyclic graphs. Also, zero clustering is obtained for all vertices in cyclic graphs having four or more vertices. The only nonzero clustering coefficients are obtained in tri-membered cycles. As seen from the last two graphs in Figure 3, at the same level of connectedness, graphs can differ considerably by the distribution of the clustering coefficients over vertices, as well as by the average degree of clustering.

It is important to note that the average clustering coefficient of biological networks is higher by 2 orders of magnitude than that of the random networks of the same size and the same average vertex degree. Thus, for the protein−protein interaction network of *Saccharomyces cerevisiae*, $<C> = 0.142$, whereas for the corresponding random network, it is only 0.001 39.[29] The clustering effects in molecular structures have not been studied, due to the fact that the highly strained tri-membered atomic rings are rarely stable. In crystallography, however, the abundance of tetrahedral structures makes cluster analysis quite relevant.

A closer correspondence between molecular structures and biological networks can be found in the recent extensions of the clustering coefficient concept, which includes larger cycles[30] as well as clustering profiles.[31]

5. DISTANCES AND CENTRALITY IN MOLECULAR GRAPHS AND BIOLOGICAL NETWORKS

The basic distance descriptors used in both fields are the same, although the terms used sometimes differ. For a graph/network having V vertices/nodes and E edges/links, distance descriptors are defined from the distance matrix **D**, the elements d_{ij} of which are integers equal to the number of edges connecting vertices i and j along the shortest path between them. Like vertex degree a_i, the *vertex* (*node) distance* d_i is obtained as a sum of all entries d_{ij} in the ith row of the corresponding graph matrix (eq 6a). However, while the vertex degree is a measure of local connectivity, the vertex distance measures the total distance to all other vertices in the graph. It is thus inversely related to vertex centrality (vide versa): the smaller the d_i, the more central the vertex. The *graph (network) distance D* is defined as the sum of all distances d_i (eq 6b), and its average value per vertex, $<d_i>$, is given by eq (6c). The Wiener index,[32] widely used in chemical graph theory[2,6] is simply half of the graph distance. The most common distance descriptor is *the average graph (network) distance*, $<d>$, or graph (network) radius (eq 7a). In the field of network studies, it is frequently named by different names, such as the average degree of separation, and even average path length. The last term is incorrect, indeed, because distances are measured only along the shortest paths, not all possible paths in the graph. Other important terms used are *vertex eccentricity*, e_i, which is the maximal vertex distance from vertex i to all other vertices, and *graph (network) diameter*, D(max), the largest distance in the entire graph.

$$d_i = \sum_{i=1}^{V} d_{ij} \tag{6a}$$

$$d(G) = \sum_{i=1}^{V} d_i = \sum_{i=1}^{V} \sum_{j=1}^{V} d_{ij} \tag{6b}$$

$$<d_i> = \frac{D}{V} \tag{6c}$$

$$<d(G)> = \frac{D}{V(V-1)} \tag{7a}$$

or

$$<d'(G)> = \frac{D}{V^2} \tag{7b}$$

The total number of distances in eq 7a is the number of nondiagonal distance matrix entries, whereas the denominator of eq 7b includes the cases with nonzero diagonal entries as well.

It was shown in section 3 above that the vertex degree distribution can be quantified by using the Shannon information function (eqs 3a and 3b). The distribution of distances over their magnitude,$\{m_i\}$, as well as that over graph nodes, $\{d_i\}$, were also used as sensitive descriptors of molecular structure:[33]

$$d_i\{d_1, d_2, d_3, \ldots, d_V\} \tag{8a}$$

$$m_i\{m(1), m(2), \ldots, m(d_{\max})\} \tag{8b}$$

$$I_{\mathrm{nd}} = D \log_2 D - \sum_{i=1}^{V} d_i \log_2 d_i \tag{9a}$$

$${}^{m}I_d = D \log_2 D - \sum_{i=1}^{d(\max)} m(i) \log_2 m(i) \tag{9b}$$

Here, $m(i)$ denotes the frequency of occurrence of the distance of magnitude i, and d_i is the total distance of node i. The two information-theoretic indices describe the total information content of the graph, defined on the node distances and on the distance magnitude, respectively. The average values of these descriptors are obtained after dividing them by the number of vertices V, and by the number of distances $V(V-1)$, respectively. For normalizing the two descriptors within the 0−1 range, eqs 9a and 9b are divided by $D \log_2 D$.

It should be mentioned that, when applied to a large-scale comparison of biological networks of different species, the information-theoretic descriptors, defined by eqs 3 and 9, and the average and normalized indices derived from them, show generally better correlation with the phylogenetic distance between the species than other single-number connectivity-based and distance-based descriptors.[34] The same conclusion (based on the comparative analysis of the networks of metabolic pathways of over 380 species)[34] was valid also when the distributions of pairs of species were used as linear sequences to compare their similarity by calculating the Jaccard coefficient. This result might be of interest for QSPR/QSAR studies in chemistry and drug design.

In directed networks, the calculation of network radius is rather complicated. Due to the inaccessibility of some vertices, the distance should be counted as an infinite one, which does not allow comparison of different directed networks. Counting such distances as zeros (wrongly used in some bioinformatics publications) leads to the absurd result that the average distance in a directed graph is smaller than that in the parent undirected ones. An improved distance measure was recently introduced, which accounts for the limited vertex accessibility in directed graphs, and enables distance-based network comparisons.[35]

The high connectivity in biological and other complex dynamic networks makes them rather compact. Yet, it came as a surprise that the radius of such huge networks is very small. This property, called *small-worldness*,[28] was confirmed for all molecular biology networks. Thus, the average degree of separation of two proteins in protein−protein interaction networks was estimated to be within the range of five to eight edges.[36] The network small size contributes greatly to the high robustness of the cell against any perturbation or attack. Molecular graphs, which are static but not dynamic, do not show this property, and the distance between two vertices in some molecular graphs could be very large.

Graph distances provide the basis for the concept of vertex *centrality*, the hierarchically ordering graph vertices, according to their location with respect to the *graph center*. The classical definition of graph center, introduced by Harary,[1] awards with a central role the vertex(es) which are the least distant from all other vertices, as measured by the *minimal eccentricity value*, e_i. This definition frequently classifies several vertices as central in molecular graphs. In the huge biological networks, the number of such central nodes is usually very large, which makes it impractical. A more detailed definition for the center of graphs was proposed in the 1980s in a hierarchical manner.[37] It assumes the Harary definition as a first criterion for centrality. When two or more vertices with different neighborhoods satisfy this criterion, a second criterion distinguishes between them by the magnitude of their *vertex distance*, d_i, the central vertices having the minimal value. When both criteria still continue to produce a group of nonequivalent central vertices, the third criterion called *vertex distance code* gives centrality precedence to vertices having a higher frequency of occurrence of distance 1, then 2, and so forth. When even the three hierarchical criteria cannot resolve the problem, an iterative algorithm for vertex and edge centrality determination completes the task.[38,39] The centric ordering of all graph vertices into groups of equal centrality containing N_1, N_2, ..., N_k vertices, respectively, has been used to define centric descriptors of graphs:[37]

$$\delta = \sum_{i=1}^{k} N_i^2 \tag{10a}$$

$$C = V \log_2 V - \sum_{i=1}^{k} N_i \log_2 N_i \tag{10b}$$

It has been shown that the centric ordering of graph vertices, calculated by eqs 10a and 10b, does not correlate with any other known vertex distribution.[40]

Several graph centrality measures have been introduced in the field of social sciences, in which the problem for determining leaders has been really central. *Closeness centrality*[41] is close in meaning to the above concept in using the condition for the vertex distance minimum as a single classifying criterion. The descriptor is defined as an inverse

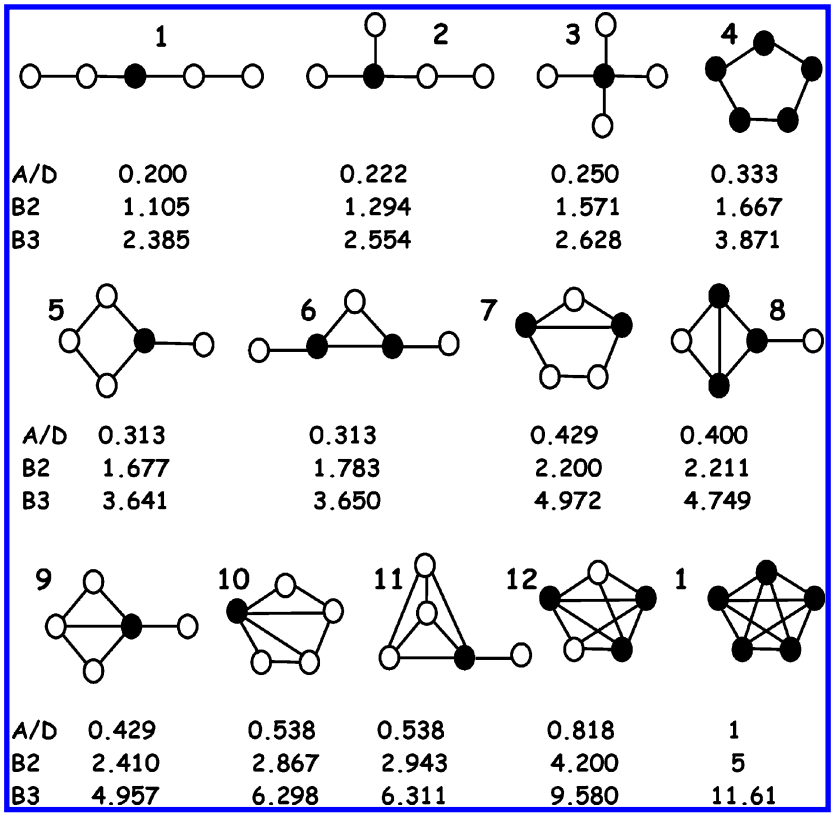

Figure 4. A total of 13 graphs with five vertices and 4−10 edges. The three indices of connected small-worldness, A/D (or B1), B2, and B3, defined by eqs 13 and 14, tend to increase from **1** to **13**. The black vertices are the corresponding graph centers as determined by the b_i vertex descriptor (eq 13b).

of the node distance d_i, normalized by dividing into the number of vertex distances ($V - 1$):

$$\text{closeness centrality} = \frac{V-1}{d_i} \tag{11}$$

Another centrality measure applied in that field, termed *betweenness centrality*,[42] attracted much attention in biological networks analysis. The idea is that it is through the central node of the network that the largest fraction of shortest paths of interaction passes through:

$$\text{betweenness centrality} = \frac{N_{\text{paths}}}{N_{\text{paths}}(\text{max})} \tag{12}$$

Other centrality measures proposed are based on normalized vertex degrees (*degree centrality*),[41] random walks,[42] the principal eigenvector and principal eigenvalue of the adjacency matrix,[43] subgraphs,[44,45] and so forth.

Centrality is of vital importance for the life of a cell. The deletion of a centrally located gene or protein is usually lethal for the cell. Such genes and proteins are called *essential*.[46] Better centrality measures have been recently introduced by Estrada[47,48] for identifying essential proteins in intracellular protein−protein interaction networks. Another group of essential genes and proteins are those which are highly connected. Node connectivity and centrality in biochemical networks were found to be of importance for longevity.[49,50] Centrality effects have not been investigated in molecules. The potential centrality effects on the reactivity of atoms and some of their physical characteristics might be expected to exist in large-size molecules.

6. COMBINING CONNECTIVITY WITH SMALL-WORLDNESS

Highly evolved biological networks are characterized by both a high average connectivity and a small radius. Therefore, a descriptor combining both trends could be a convenient measure of network evolution and complexity.[27] Such descriptors are introduced here in three versions. The simplest one is just the ratio of total adjacency A and total distance D, denoted as the A/D or B1 index (eq 13a). Using single numbers to characterize an entire network is always associated with a loss of information. Hence, the A/D index can be used mainly for fast estimates of network complexity. Considerably better results can be obtained proceeding from the entire vertex degree and vertex distance distributions. One way to proceed from here is to calculate for each vertex the ratio of its degree a_i and distance d_i. This results in a new vertex invariant b_i (eq 13b), which in turn can be considered as a simple vertex centric index, somewhat similar to closeness centrality (eq 11). After summing up the b_i values over all vertices, one arrives at the B2 descriptor (eq 13c), which is much more sensitive to subtle details of network topology. This is evidenced by the considerably lesser degeneracy (the same descriptor value for different graphs) and the better correlation with the ordering of graphs according to their increasing complexity (Figure 4).

$$\text{B1} = \frac{A}{D} \tag{13a}$$

$$b_i = \frac{a_i}{d_i} \tag{13b}$$

$$B2 = \sum_{i=1}^{V} b_i = \sum_{i=1}^{V} \frac{a_i}{d_i} \qquad (13c)$$

Another way to transform a distribution into a single number, capturing well the essential structural information, is to apply the Shannon information equation,[3,4] as shown in eq 14:

$$B3 = B2 \log_2 B2 - \sum_{i=1}^{V} b_i \log_2 b_i \qquad (14)$$

As seen from Figure 4, all three B indices increase with the appearance of an additional branch or the formation of a new cycle. At the same number of cycles, the three indices differ in their behavior. The B2 index shows a systematic pattern of favoring the appearance of a higher degree in one vertex and a lower degree in another vertex, rather than two equal degrees, such as 3,1 versus 2,2 and 4,2 versus 3,3. The *A*/*D* index shows degeneracy in two pairs of graphs (5/6, and 10/11). B2 and B3 indices show no degeneracy and discriminate all 13 five-vertex graphs.

The central vertices of the 13 graphs, determined by the b_i index (eq 13b), are shown in Figure 4 in black. Due to the small size of the graphs examined, the distances do not play a big role, and the centrality of vertices is dictated by their degree. The situation is totally different in the large biological networks, in which distances play a much more important role.

7. SUBGRAPHS AS MOTIFS, MODULES, AND CLIQUES

The role of specific parts of the molecular skeleton was recognized early in chemistry and expressed in the concept of atomic groups, which conserve their functions in different atomic environments. This structural chemistry concept was first formalized by Smolenski[51] in 1964, by considering different subgraphs of a molecular graph. The idea was developed in detail by Gordon and Kennedy[52] in 1973 and widely applied in the mid 1970s for QSPR/QSAR within the framework of the molecular connectivity concept of Kier ad Hall.[53]

The latest development in post-genome biology rediscovered the importance of graph fragments of different sizes. The smallest pieces of biological networks, analyzed in the laboratory of Alon et al., were called *motifs*.[54] However, only those of the smallest subgraphs are motifs, which are considerably more abundant in the network than in the randomized networks having the same number of nodes and links. As shown in subsequent publications of this group,[55] some of the smallest motifs play an essential role in gene regulatory networks. The list of motifs' abundances is used as fingerprints for different types of networks in different species. An example of such simple motifs, along with their names, is shown in Figure 5.

Considerably larger subgraphs are used in metabolic, protein−protein interaction and other biological networks as *modules*,[18,19,56] which perform certain biological functions. Such modules are frequently called *pathways*. Typical examples are the glycolysis metabolic pathway, which is the major source of energy supply for the organism, and the apoptosis signaling pathway, which executes cell suicide in a variety of cases, including fighting cancer. Mathematically, a module is defined as a group of nodes which interact more between themselves rather than with the remaining network nodes. However, there is no direct correspondence between structural and functional modules, and the search for finding the modularization technique that would best correspond to functional modules continues. Traditional chemical research rarely includes large molecular fragments, except maybe in the case of benzenoid polycyclic hydrocarbons. The development of nanotechnology is rapidly changing the scene, and nanomodules emerge as a good chemical analog to the modules of cellular networks.

The maximal subgraphs, called *cliques*, are also of interest in biological networks. These are complete subgraphs, every node in which is connected to every other one (Figure 6).

Clique properties of networks have been widely applied in computer sciences but so far have eluded the attention of theoretical chemists. They might be of interest in the topological analysis of crystals and atomic clusters containing tetrahedral cells.

8. COMPLEXITY OF GRAPHS AND NETWORKS

The complexity of graphs has been a subject of intensive studies in chemical graph theory since the beginning of the 1980s.[57−60] The initial approaches have been based on a variety of single-number descriptors. The more mature concepts that appeared in the second half of the 1990s have been based on a more detailed representation of topological structure using the totality of subgraphs or walks in the graph.[61−64] Besides the total count of subgraphs, SC, and walks, WC, distributions of subgraphs of different size, ^{e}SC (e = number of edges), and distributions of walks of different length l, ^{l}WC, have been used for a more adequate representation of graph complexity. One more conceptual step in the quantitative assessment of complexity has been made by accounting for the fact that different subgraphs also have different complexity. The simplest measure of the subgraph complexity is their total adjacency (eq 1b). Combining the concept of subgraphs and connectivity resulted in an effective complexity measure, called *overall connectivity*, OC. The partitioning of the subgraph count, overall connectivity, and walk count into their different size

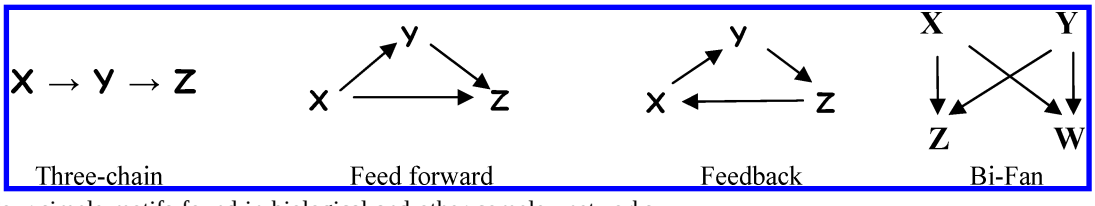

Figure 5. Four simple motifs found in biological and other complex networks.

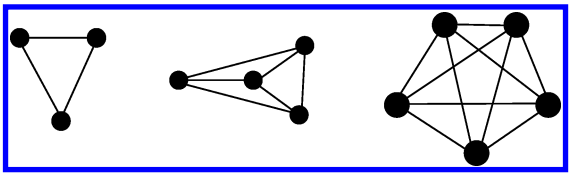

Figure 6. Examples of complete subgraphs (cliques) with 3, 4, and 5 vertices.

constituents is shown in eqs 15a−c:

$$SC = {}^{0}SC + {}^{1}SC + {}^{2}SC + ... + {}^{E}SC \quad (15a)$$

$$OC = {}^{0}OC + {}^{1}OC + {}^{2}OC + ... + {}^{E}OC \quad (15b)$$

$$WC = {}^{1}WC + {}^{2}WC + {}^{3}SC + ... + {}^{V-1}WC \quad (15c)$$

Complexity measures 13a−c have been shown to capture the patterns of increasing graph complexity with the increase in the number of cycles, branches, vertices of higher degree, and so forth. They have been applied after some adaptation as the first measures of biological network complexity, along with the information-theoretic index for the vertex degree distribution (eqs 3a,b). Due to the huge size of biological networks, which leads to combinatorial explosion in the number of subgraphs and walks, we recommended to use for network complexity estimates only the first three terms in eqs 13a−c.

9. CONCLUSION

This study analyzed the interplay between the methods used to characterize molecular structure in mathematical and computational chemistry and those used in mathematical and computational biology to characterize biological networks. Both areas heavily rely on applied graph theory. The expertise accumulated during the 40 years of chemical graph theory proved to be of use in characterizing network topology, offering the first measures of network complexity, and new measures of network centrality. Also fruitful was the transfer of expertise in introducing information-theoretic descriptors based on graph invariants. Besides the variety of information indices based on connectivity and distances in graphs, new such measures have been introduced for characterizing the distribution of cliques in networks. Also, contribution from chemical graph theory to a biological one was a result of the development of a practical realistic measure for the average distance in directed networks, avoiding the difficulties caused by the inaccessibility of some vertices in these networks.

Was there any feedback from biological networks to chemical graph theory? The answer is definitely, "yes". Biology has attracted many prominent researchers from other areas of applied graph theory. Their expertise in network motifs, modules, and cliques can stimulate applications in different areas of theoretical chemistry, for example, dealing with mineralogy and crystallography, atomic clusters, and nanotechnology. The concept of centrality, which has been only marginally studied in molecular structures, was highly developed and found of great interest for biological networks, which could be incapacitated by eliminating a central node, an event leading to cell death. These ideas might prove of interest for some specific areas of organic chemistry. Examples along this line are already starting to appear in the literature, such as applying the bipartivity measure of networks[65] to the stability of fullerenes[66] and branching of the molecular skeleton.[67] Chemical graph theory may use for a variety of purposes the new B indices, introduced as an integral characteristic of connectivity and small-world properties of biological networks. QSAR/QSPR studies and data mining in drug design may benefit from investigating in more detail the conclusions for a potential advantage of direct use of distributions of graph invariants or their transformation into information-theoretic indices, rather than some of the common topological descriptors. Last but not least, the lessons from dynamic biological networks might stimulate more interest in developing a dynamic chemical graph theory, looking beyond the static molecular graphs.

REFERENCES AND NOTES

(1) Harary, F. *Graph Theory*, 2nd ed.; Addison-Wesley: Reading, MA, 1969.
(2) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.
(3) Shannon, C.; Weaver, W. *Mathematical Theory of Communications*; University of Illinois Press: Urbana, IL, 1949.
(4) Bonchev, D. *Information-Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, U. K., 1983.
(5) *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Amsterdam, 1999.
(6) *Topology in Chemistry. Discrete Mathematics of Molecules*; Rouvray, D. H., King, R. B., Eds.; Horwood: Chichester, U. K., 2002.
(7) Barabási, A.-L.; *Linked. The New Science of Networks*; Perseus: Cambridge, MA, 2002.
(8) Dorogovtsev, S. N.; Ferreira Mendes, J. F.; Ioffe, A. F. *Evolution of Networks. From Biological Nets to the Internet and WWW*; Oxford University Press: Oxford, U. K., 2003.
(9) Newman, M.; Barabási, A.-L.; Watts, D. J. *The Structure and Dynamics of Networks*; Princeton University Press: Princeton, NJ, 2006.
(10) Watts, D. J. *Small Worlds: The Dynamics of Networks between Order and Randomness*; Princeton Universtity Press: Princeton, NJ, 1999.
(11) Barabási, A.-L.; Albert, R. Emergence of Scaling in Random Networks. *Science (Washington, DC, U.S.)* **1999**, *286*, 509−512.
(12) Barabasi, L.-A.; Oltvaj, L. Z. Network Biology: Understanding the Cell's Functional Organization. *Nat. Genet.* **2004**, *5*, 102−114.
(13) Kitano, H. Computational Systems Biology. *Nature (London, U.K.)* **2002**, *420*, 206−210.
(14) Sharan, R.; Suthram, S.; Kelley, R. M.; Kuhn, T.; McCuine, S.; Uetz, P.; Sittler, T.; Karp, R. M.; Ideker, T. From the Cover: Conserved Patterns of Protein Interaction in Multiple Species. *PNAS* **2005**, *102*, 1974−1979.
(15) Sharan, R.; Ideker, T. Modeling Cellular Machinery through Biological Network Comparison. *Nat. Biotechnol.* **2006**, *24*, 427−433.
(16) *Pathway Studio*, version 4.0; Ariadne Genomics Inc.: Rockville, MD, 2006.
(17) Yamada, T.; Goto, S.; Kanehisa, M. Extraction of Phylogenetic Network Modules from Prokaryote Metabolic Pathways. *Genome Inf. Ser.* **2004**, *15*, 249−258.
(18) Papin, J. A.; Reed, J. L.; Palsson, B. O. Hierarchical Thinking in Network Biology: The Unbiased Modularization of Biochemical Networks. *Trends Biochem. Sci.* **2004**, *29*, 641−647.
(19) Newman, M. E. J. Modularity and Community Structure in Networks. *PNAS* **2006**, *103*, 8577−8582.
(20) Bonchev, D. Complexity Analysis of Yeast Proteome Network. *Chem. Biodiversity* **2004**, *1*, 312−326.
(21) Nitta, K.-H. A Graph-Theoretical Approach to Statistics and Dynamics of Tree-Like Molecules. *J. Math. Chem.* **1999**, *25*, 133−143.
(22) Bonchev, D.; Markel, E.; Dekmezian, A. Long-Chain Branch Polymer Dimensions: Application of Topology to the Zimm−Stockmayer Model. *Polymer* **2002**, *43*, 203-222.
(23) Gutman, I.; Toropov, A. A.; Toropova, A. P. The Graph of Atomic Orbitals and Its Basic Properties 1. Wiener Index. *MATCH* **2005**, *53*, 215−224.
(24) Galvez, J.; Garcia-Domenech, R.; de Julian-Ortiz, J. V. Assigning Wave Functions to Graphs: A Way to Introduce Novel Topological Indices. *MATCH* **2006**, *56*, 509−518.
(25) Bonchev, D. Complexity of Protein−Protein Interaction Networks, Complexes and Pathways. In *Handbook of Proteomics Methods*; Conn, M., Ed.; Humana: New York, 2003; pp 451−462.

(26) Bonchev, D. Shannon's Information and Complexity. In *Mathematical Chemistry Series*, "*Complexity in Chemistry*"; Bonchev, D., Rouvray, D. H., Eds.; Taylor & Francis: London, U. K., 2003; Vol. 7, pp 155−187.
(27) Bonchev, D.; Buck, G. A. Quantitative Measures of Network Complexity In *Complexity in Chemistry, Biology and Ecology*; Bonchev, D., Rouvray, D. H., Eds.; Springer: New York, 2005; pp 191−235.
(28) Watts, D. J.; Strogatz, S. H. Collective Dynamics of "Small-World" Networks. *Nature* **1998**, *393*, 440−442.
(29) Fernandez, P.; Solé, R. V.; Graphs as Models of Large-Scale Biochemical Organization. In *Complexity in Chemistry, Biology and Ecology*; Bonchev, D., Rouvray, D. H., Eds.; Springer: New York, 2005; pp 155−189.
(30) Fronczak, A.; Holyst, J. A.; Jedynak, M.; Sienkiewicz, J. Higher Order Clustering Coefficients in Barabaśi−Albert Networks. *Physica A* **2002**, *316*, 688−694.
(31) Abdo, A. H.; de Moura, A. P. S. *Clustering as a Measure of the Local Topology*; arXiv: physics/0605235; arXiv.org ePrint archive, 2006. http://arxiv.org/physics/(0605235) (accessed Sept 14, 2006).
(32) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(33) Bonchev, D.; Trinajstic, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517−4533.
(34) Mazurie, A.; Bonchev, D.; Buck, G. A. Networks of Interacting Pathways Reflect Phylogenetic Relationships. *Nat. Genet.* (submitted for publication).
(35) Bonchev, D. On the Complexity of Directed Biological Networks. *SAR QSAR Environ. Res.* **2003**, *14*, 199−214.
(36) Yook, S.-H.; Oltvai, Z. N.; Barabási, A.-L. Functional and Topological Characterization of Protein Interaction Networks. *Proteomics* **2004**, *4*, 928−942.
(37) Bonchev, D.; Balaban, A. T.; Mekenyan, O. Generalization of the Graph Center Concept, and Derived Topological Indexes. *J. Chem. Inf. Comput. Sci*. **1980**, *20*, 106 −113.
(38) Bonchev, D.; Mekenyan, O.; Balaban, A. T. An Iterative Procedure for the Generalized Graph Center in Polycyclic Graphs. *J. Chem. Inf. Comput. Sci*. **1989**, *29*, 91−97.
(39) Bonchev, D. The Concept for the Center of a Chemical Structure and Its Applications. *THEOCHEM* **1989**, *185*, 155−168.
(40) Motoc, I.; Balaban, A. T.; Mekenyan, O.; Bonchev, D. Topological Indices: Inter-Relations and Composition. *MATCH* **1982**, *13*, 369−404.
(41) Freeman, L. C. Centrality in Social Networks: Conceptual Clarification. *Social Networks* **1979**, *1*, 215−239.
(42) Newman, M. E. J. A Measure of Betweenness Centrality Based on Random Walks. *Social Networks* **2007** (submitted).
(43) Bonacich, P. Factoring and Weighting Approaches to Status Scores and Clique Identification. *J. Math. Sociol.* **1972**, *2*, 113−120.
(44) Estrada, E.; Rodríguez-Velázquez, J. A. Subgraph Centrality in Complex Networks. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2005**, *71*, 056103.
(45) Estrada, E.; Rodríguez-Velázquez, J. A. Subgraph Centrality and Clustering in Complex Hyper-Networks. *Physica A* **2006**, *364*, 581−594.
(46) Gavin, A. C.; Bösche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A.-M.; Cruciat, C.-M.; Remor, M.; Höfert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M.-A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B., Neubauer, G.; Superti-Furga, G. Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. *Nature (London, U.K.)* **2001**, *411*, 41−42.
(47) Ernesto, E. Protein Bipartivity and Essentiality in the Yeast Protein−Protein Interaction Network. *J. Proteome Res.* **2006**, *5*, 2177−2184.
(48) Estrada, E. Virtual Identification of Essential Proteins within the Protein Interaction Network of Yeast. *Proteomics* **2006**, *6*, 35−40.
(49) Promislow, D. E. L. Protein Networks, Pleiotropy and the Evolution of Senescence. *Proc. R. Soc. Edinburgh, Sect. B: Biol. Sci.* **2004**, *271*, 1225−1234.
(50) Witten, T. M.; Bonchev, D. Longevity Gene Network Analysis for the Nematode *C. elegans*. *Chem. Biodiversity* (Submitted).
(51) Smolenski, E. A. Graph-Theory Application to the Calculations of Structural-Additive Properties of Hydrocarbons. *Zh. Fiz. Khim.* **1964**, *38*, 1288−1291.
(52) Gordon, M.; Kennedy, J. W. The Graph-Like State of Matter. Part 2. LCGI Schemes for the Thermodynamics of Alkanes and the Theory of Inductive Inference. *J. Chem. Soc., Faraday Trans. 2* **1973**, *69*, 484−504.
(53) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(54) Milo, R.; Itzkovitz, S.; Kashtan, N.; Levitt, R.; Alon, U. Network Motifs: Simple Building Blocks of Complex Networks. *Science (Washington, DC, U.S.)* **2002**, *298*, 824−827.
(55) Shen-Orr, S.; Milo, R.; Mangan, S.; Alon, U. Network Motifs in the Transcriptional Regulation Network of *Escherichia coli*. *Nat. Genet.* **2002**, *31*, 64−68.
(56) Ravasz, E.; Somera, A. L.; Mongru, D. A.; Oltvai, Z. N.; Barabaśi, A.-L. Hierarchical Organization of Modularity in Metabolic Networks. *Science (Washington, DC, U. S.)* **2002**, *297*, 1551−1555.
(57) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.
(58) Bonchev, D.; Polansky, O. E. On the Topological Complexity of Chemical Systems. In *Graph Theory and Topology in Chemistry*; King, R. B., Rouvray, D. H., Eds.; Elsevier: Amsterdam, 1987; pp 126−158.
(59) *Mathematical Chemistry Series*; Bonchev, D., Rouvray, D. H., Eds.; Taylor & Francis: London, U. K., 2003; Vol. 7 "Complexity in Chemistry".
(60) *Complexity in Chemistry, Biology and Ecology*; Bonchev, D., Rouvray, D. H., Eds.; Springer: New York, 2005.
(61) Bertz, S. H.; Sommer, T. J. Rigorous Mathematical Approaches to Strategic Bonds and Synthetic Analysis Based on Conceptually Simple New Complexity Indices. *Chem. Commun.* **1997**, 2409−2410.
(62) Bonchev, D. Novel Indices for the Topological Complexity of Molecules. *SAR QSAR Environ. Res.* **1997**, *7*, 23−43.
(63) Rücker, G.; Rücker, C. Substructure, Subgraph and Walk Counts as Measures of the Complexity of Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1457−1462
(64) Nikolić, S.; Trinajstić, N.; Tolić, I. M.; Rücker, G.; Rücker, C. On Molecular Complexity Indices. In *Mathematical Chemistry Series*; Bonchev, D., Rouvray, D. H., Eds.; Taylor & Francis: London, U. K., 2003; Vol. 7 "Complexity in Chemistry"; pp 29−89.
(65) Estrada, E.; Rodríguez-Velázquez, J. A. Spectral Measures of Bipartitivity in Complex Networks. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2005**, *72*, 055510.
(66) Došlić, T. Bipartivity of Fullerene Graphs and Fullerene Stability. *Chem. Phys. Lett.* **2005**, *412*, 336−340.
(67) Estrada, E.; Rodriguez-Velazquez, J. A.; Randic, M. Atomic Branching in Molecules. *Int. J. Quantum Chem.* **2006**, *106*, 823−832.

CI7000617