

## Complexity Analysis of Yeast Proteome Network<sup>1)</sup>

by **Danail Bonchev**

Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284-2030, USA (e-mail: dgbonchev@mail1.vcu.edu)

and

Program for Theory of Complex Natural Systems, Texas A&M University, Galveston, TX 77551, USA (e-mail: bonchevd@sbcglobal.net)

---

Topological and compositional complexity of protein–protein networks is assessed in a variety of ways making use of graph theory and information theory. The methodology used is borrowed from mathematical chemistry and includes complexity descriptors such as substructure count, overall connectivity, walk count, and information on various vertex distributions. The approach is applied to the (incomplete) proteome of *Saccharomyces cerevisiae* containing 232 protein complexes of a total of 1,440 proteins. The proteome network and each of its nine functional subsets of protein complexes are disconnected graphs, containing a number of noninteracting species and a major component. A weighted edge between two vertices in these graphs stands for the number of shared proteins between the respective complexes. The major component is a highly connected, ‘small-world’ network, in which the average vertex distance between protein complexes does not exceed 2.2 (2.4 for the entire proteome), whereas the maximum distance does not exceed 4 (or 5 for the proteome). The vertex degree distribution in the major proteome component with 199 complexes follows the power law  $P(k) \sim k^{-\gamma}$ , with  $\gamma \approx 1.7$ . The analysis of the functional organization of the yeast proteome has shown that, for any pair of biological functions, there always exist many proteins that can perform both functions. The potential application of the quantitative proteome descriptors discussed includes quantitative relationships between the structure and biological action of dynamic protein complexes in changing environment, identification of targets for markers/drugs, as well as system analysis and comparative studies of proteomes.

---

**1. Introduction.** – Recent advances in the strategies and techniques for characterization of multiprotein complexes have made feasible the build-up of large-scale networks of protein interactions [1–7]. This offered for the first time the chance to analyze, on a molecular level, the work of the cellular machinery in its entirety. The early stages of proteome research dwelled predominantly on protein over- and underexpression, but not on protein–protein interaction, thus touching only the tip of the iceberg. However, one might expect more-definite conclusions on the role of a certain protein to be drawn only after knowing its specific location in the biological pathway of interest, as well as in the global protein network, because the individual pathways are interdependent. Containing fundamental biological information, the large-scale networks of functionally linked proteins provide the basis for a variety of applications. Mathematical analysis of these networks could contribute not only to the better understanding of the biological machine; it provides tools for extracting valuable information, establishing similarities, finding characteristic patterns, and deriving quantitative structure/activity relationships, thus contributing to the identification and

---

<sup>1)</sup> Presented, in part, at *IBC's Annual Proteomics 2002 Conference*, Philadelphia, May 6–9.

screening of potential drug or marker candidates. To this end, the present study introduces quantitative measures of the complexity of proteome and protein complexes by extending previous extensive complexity analysis of molecular structure [8][9] and chemical reaction networks [10]. Differing from the quantitative characterizations of 2D and 3D electrophoretic proteomic maps [11][12], our approach focuses on large-scale protein–protein-interaction maps.

**2. Complexity of Structures and Networks.** – The concept of *complexity* as a general property of every system, independent of the nature of the system, began crystallizing in the 1980s, although its origin might be traced back to the 1950s, *i.e.*, to the pioneering work of *Rashewsky* [13]. Complexity is a multifaceted notion, often being defined in a hierarchical manner. In this paper, our analysis deals with *structural complexity*, the kind of complexity that accounts for the manner in which the elements of a system are organized (or connected) in a structure. We also define *compositional complexity*, which reflects the distribution of the structure elements into different classes, depending on their physical nature or other partitioning criteria. The methods used – graph theory [14] and information theory [15] – are general; they do not depend on the specific nature of the system's elements. The topological and information descriptors thus defined might be applied to a wide range of systems and phenomena. In what follows, they will be defined specifically for protein networks regarded as graphs, *i.e.*, structures built by vertices connected by edges. The vertex in such a graph may represent an individual protein, a protein complex, or even a group of protein complexes. Both undirected and directed graphs can be used, depending on the information available on the individual protein–protein interactions.

**2.1 Topological Complexity Descriptors.** The number of vertices, edges, and cycles are the primary descriptors of graph topology. The number of graph vertices  $V$  characterizes the structure size. The connectivity of the structure elements is expressed by the number of graph edges  $E$ . An essential component of topological complexity is the number of cycles  $C$  it contains. The more connected the structure, the more edges and cycles it contains. These fundamental descriptors of structure connectivity are interrelated by the *Euler* equation:

$$C = E - V + 1 \quad (1)$$

The *connectivity* of protein networks,  $Cn = 100(V_{conn}/V)$ , has been assessed [3] as the percentage of the vertices that are connected with at least one other vertex. In a more-precise way, network connectivity is defined by *connectedness* (or *connectance*),  $Conn$ , which takes into account *all* connections  $E$  in the network:

$$Conn[\%] = \frac{2E}{V(V-1)} \times 100 \quad (2)$$

Here,  $V(V-1)/2$  is the maximum possible number of edges in a simple graph having  $V$  vertices. Thus, by definition,  $Conn$  is normalized to have values between 0 and 100% for a completely disconnected or a completely connected (or complete) graph,

respectively. In multigraphs, more than one edge can exist between two vertices, and *Conn* can exceed 100%. However, it makes sense even for a network characterized by a multigraph to calculate not only this *multiple connectedness*, but also the *basic connectedness*, which counts only the presence or absence of connection between any two vertices but not the number of these pairwise connections.

The local connectivity of a vertex  $i$  in the network is described by  $a_i$ , the *vertex degree*, which counts the number of the vertex-nearest neighbors. Again, when multiple edges are present, one may define the *multiple vertex degree* as the sum of all edges emanating from the vertex. The sum of the vertex degrees,  $A = \sum a_i$ , is termed *total adjacency*. The average vertex degree  $\langle a_i \rangle = A/V$  is also used as an average measure for the network connectivity. In directed graphs, edges have a direction. As a result, one distinguishes ‘in-degree’ and ‘out-degree’ of each vertex. Proteins (or protein complexes) with a high in-degree are often end points of biological pathways and could be of interest as potential targets for drug design.

Other essential topological characteristics of a network are the distances between the network nodes. The distance  $d_{ij}$  between the vertices  $j$  and  $i$  is an integer quantity equal to the number of edges connecting the vertices along the shortest path between them. The *vertex distance*,  $d_i$ , termed also *distance degree*, is the sum of the distances between the vertex  $i$  and all other vertices in the network. Network distance, widely known in mathematical chemistry as the *Wiener number*  $W$  [16][17], is the sum of all distances in the network:

$$W = \frac{1}{2} \sum_{i,j=1}^V d_{ij} = \frac{1}{2} \sum_{i=1}^V d_i \quad (3)$$

The *average vertex distance*,  $\langle d_i \rangle = 2W/V$ , characterizes how easily a vertex can reach all other vertices. A useful parameter is also the *average intersite distance* or *graph radius*,  $\langle d \rangle = 2W/V(V-1)$ , which shows how easily one vertex can reach another vertex. *Graph diameter* is usually defined as the largest intersite distance in the graph. Biological networks are characterized as ‘small-world networks’ [18]; the average distance in them is small [19]. In networks with directed edges, there could be no paths connecting some pairs of vertices. Strictly speaking, the distance between such a pair of vertices is infinite, which makes it impossible to define the *Wiener number*. However, for practical purposes, one may calculate distance descriptors by counting only the distances along the existing paths in the network [20][21].

The topological descriptors described in the foregoing are simple and easy to calculate. However, they often do not satisfy the major criterion to be nondegenerate, *i.e.*, to provide a different value for each individual structure. For this reason, sophisticated hierarchically built complexity measures have been developed [9] to better distinguish among structures. For the large protein–protein networks, we offer to use only the first terms in such measures; for smaller networks, higher-order terms can similarly be applied.

A modern concept of structural complexity regards a structure to be more complex when it contains more substructures [22–24]. One can, thus, evaluate network

complexity by the number of subgraphs  $SC$  (*subgraph count*) it contains, beginning with the vertices  $V = {}^0SC$ , edges  $E = {}^1SC$ , two-edge subgraphs (the *Platt* index [25], applied first by *Bertz* [26] as a complexity measure)  $Pl = {}^2SC$ , etc.:

$$SC = {}^0SC + {}^1SC + {}^2SC + \dots + {}^E SC \quad (4)$$

Subgraph count is a sensitive complexity measure, however, for large-scale networks, the computational time required might be considerable. For such cases, we propose to use instead the second- and third-order-subgraph counts,  ${}^2SC$  and  ${}^3SC$ , respectively, or the respective partial sums,  ${}^{0-2}SC$  and  ${}^{0-3}SC$ .

Another hierarchically built complexity concept is that of *overall topological indices* [27]. It extends the idea of using all substructures by summing up the values of some simple graph invariants for each substructure. When the invariant selected is the vertex degree, the resulting complexity measure is called *overall connectivity* [28],  $OC$ . Therefore, the higher the connectivity of the graph and all of its subgraphs, the higher the graph complexity. When the subgraph is characterized by its *Wiener* number, one defines the *overall Wiener index* [29],  $OW$ , etc. The overall connectivity was shown to increase with all complexifying factors, such as size, branching, cyclicity, and centrality, as well as to produce excellent structure/property correlations with physicochemical properties of chemical compounds [27]. We recommend for complexity assessments of large networks to use the initial terms  ${}^1OC$  and  ${}^2OC$  in the expression for the overall connectivity or their sum  ${}^{0-2}OC$ :

$$OC = {}^0OC + {}^1OC + {}^2OC + \dots + {}^E OC \quad (5)$$

$${}^1OC = \sum_{i,j\text{-adjacent}} (a_i + a_j); {}^2OC = \sum_{i-j-k\text{-adjacent}} (a_i + a_j + a_k) \quad (6)$$

*Rücker* and *Rücker* [30][31] proposed to measure the complexity of a graph by the so-called *total walk count*,  $twc$ . (A *walk* is a path between two vertices that can repeatedly traverse vertices and/or edges.) The idea is similar to that of the subgraph count: the more walks in the graph, the more complex it is. The total walk count can also be presented as a sum of hierarchically ordered terms of first, second, etc. order, which count the walks of length one, two, etc., respectively. The first several terms, such as  ${}^2wc$ ,  ${}^3wc$ , and their partial sums  ${}^{0-2}wc$  or  ${}^{0-3}wc$  can also be used as complexity descriptors of large protein networks.

**2.2 Information-Theoretic Descriptors.** Consider a network composed of  $N$  elements distributed according to a certain equivalence criterion  $\alpha$  into  $k$  sets, having  $N_i$  elements each. We will now define the *compositional complexity* of the network. Networks with even distribution of their elements onto the  $k$  subsets may be regarded as less complex than those with uneven distribution. Quantitatively, this complexity aspect is characterized by *Shannon's* information theory [15]. The latter ascribes the maximum *entropy of information*,  $H(\alpha)$ , to the even distribution, and the deviation of the system entropy from its maximum value defines the *information content*,  $I(\alpha)$ , of the system:

$$I(\alpha) = H_{\max}(\alpha) - H(\alpha); H(\alpha) = N \log_2 N - \sum_{i=1}^k N_i \log_2 N_i, \quad (7)$$

when  $N_1 = N_2 = \dots = N_k = 1$ , one obtains:

$$H_{\max}(\alpha) = N \log_2 N, \quad (8)$$

with

$$I(\alpha)[bits] = \sum_{i=1}^k N_i \log_2 N_i \quad (9)$$

In *Eqns. 7–9*, the logarithm at base two is used to quantify information in binary units (bits). One may also normalize *Eqn. 9* by dividing by  $H_{\max}$ , so as to define the *relative information content*,  $I_r(\alpha)$  within the range 0 to 1.

$$I_r(\alpha) = \frac{1}{N \log_2 N} \sum_{i=1}^k N_i \log_2 N_i \quad (10)$$

We may now apply this formula to define the compositional complexity of the proteome *via* its normalized information content, regarding the distribution of protein complexes into  $k$  groups according to their biological function,  $I_{func}(\text{proteome})$ , or to the localization of protein complexes within the cell,  $I_{loc}(\text{proteome})$ , respectively. At the lower hierarchical level, one may similarly calculate the normalized information content of each functional or localized set of complexes,  $I(\text{func. set})$  and  $I(\text{local. set})$ , regarding the distribution of  $N$  proteins into  $k$  complexes. Similarly, when one regards the distribution of protein complexes (or proteins) onto components (disconnected subgraphs),  $I(\text{components})$  is defined. Other information-theoretic indices have also been advanced and used for characterization of chemical structures [32–34].

The total complexity of a given proteome can be defined in a hierarchical manner (see *Fig. 1*) for the complexity descriptor  $D$ , which can be any of the topological or information theoretic indices introduced in the foregoing:

$$D(\text{proteome}) = D_{func}(\text{proteome}) + \sum_{i=1}^{i(\max)} D_i(\text{func. set } i) + \sum_{i=1}^{i(\max)} \sum_{j=1}^k D_{ij}(\text{complex } j) \quad (11)$$

or alternatively:

$$D(\text{proteome}) = D_{loc}(\text{proteome}) + \sum_{i=1}^{i(\max)} D_i(\text{loc. set } i) + \sum_{i=1}^{i(\max)} \sum_{j=1}^k D_{ij}(\text{complex } j) \quad (12)$$

The third term in the above two equations enables the complexity of each of the protein complexes to be evaluated, provided the connectivity of proteins in the complex is known.

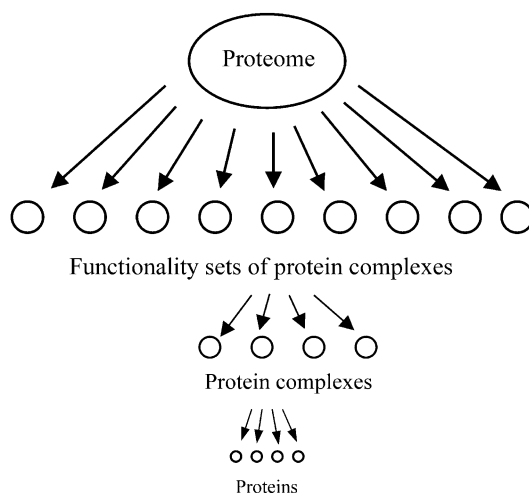


Fig. 1. The proteome hierarchical organization

Information-theoretic formalism can be applied to characterize not only the compositional complexity of the network, but also its topological complexity. The basic topological distributions of network elements are those of vertex degrees  $a_i$  and vertex-distance degrees  $d_i$ :  $\{N_1(a_1), N_2(a_2), \dots, N_k(a_k)\}$  and  $\{N_1(d_1), N_2(d_2), \dots, N_k(d_k)\}$ , where  $N_i(a_i)$  and  $N_i(d_i)$  stand for the number of vertices with vertex degree  $a_i$  and vertex-distance degree  $d_i$ , respectively. Thus, the information indices calculated for these two distributions by Eqns. 9 and 10 would reflect the equivalence or symmetry of the network vertices. However, symmetry is generally regarded as a simplifying rather than a complicating factor. The most-symmetric structure with its entire set of elements equivalent has zero information content. An alternative way of applying information theory to discrete systems has been proposed by *Bonchev* and *Trinajstić* [35]. Instead of dealing with the *number* of structure elements and its distribution in equivalence classes, this approach regards  $N$  in Eqns. 7–10 as a certain *property* or *weight* or *magnitude*, characterizing the entire system,  $N_i$  being the respective value for the  $i$ th element. Introducing in Eqns. 7–10 the sum of vertex degrees ( $vd$ ) as *total adjacency*,  $A = \sum a_i$ , and the sum of all vertex-distance degrees ( $dd$ ) as  $D = \sum d_i = \sum n_i d(i)$ , where  $d_i = 1, 2, \dots, d_{\max}$ , one obtains for the relative information on these distributions in a network with  $V$  vertices:

$$I_r(vd) = \frac{1}{A \log_2 A} \sum_{i=1}^V a_i \log_2 a_i \quad (13)$$

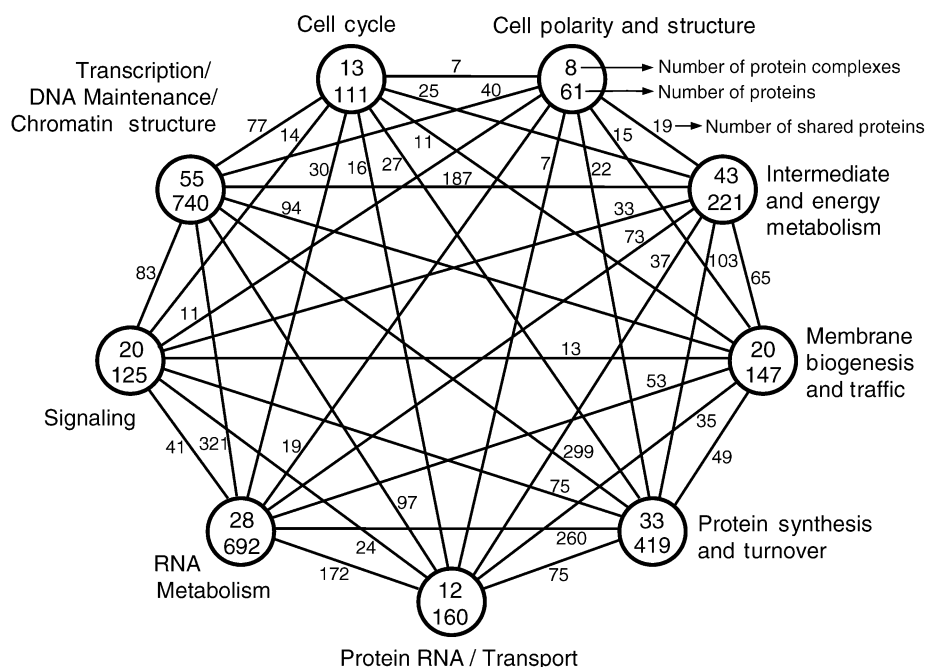
$$I_r(dd) = \frac{1}{D \log_2 D} \sum_{i=1}^V d_i \log_2 d_i \quad (14)$$

$$I_r(d) = \frac{1}{D \log_2 D} \sum_{i=1}^{d(\max)} n_i d(i) \log_2 d(i). \quad (15)$$

In what follows, we shall use *Eqn. 10* and *Eqns. 13–15* for the calculation of information-theoretic indices. Thereby, the subscript ‘*r*’ will be omitted for brevity.

**3. Analysis of the Yeast Proteome.** – 3.1 *Functional and Localization Proteome Complexity.* The methodology for assessment of proteome complexity was applied to the data on yeast proteome (*Saccharomyces cerevisiae*) published by *Gavin et al.* [1]. The data set contains the interactions of 1,440 distinct proteins (*ca.* 25% of the entire proteome), which form 232 complexes. The complexes are regarded organized in nine groups of different biological function (*Fig. 2*): cell cycle; cell polarity and structure; intermediate and energy metabolism; membrane biogenesis and traffic; protein synthesis and turnover; protein RNA and transport; RNA metabolism; signaling; and transcription, DNA maintenance, chromatin structure. They are also regarded as being distributed in different subcellular compartments: nucleus; cytoplasm; membrane; mitochondria; and ER, *Golgi* apparatus, vesicles.

In *Fig. 2*, the functional organization of the *Saccharomyces cerevisiae* proteome and the corresponding complexity descriptors are shown. The proteome is depicted as an undirected graph, each of the vertices of which represents the proteins and their



Topological descriptors:  $V = 9$ ,  $E = 36$ ,  $C = 28$ ,  $Conn = 100\%$ ,  $\langle d \rangle = 1$ ,  $a_i = d_i = 8$ ,  $\langle a_i(\text{multi}) \rangle = 562$ ,  $W = 36$ ,  $Pl = 252$ ,  ${}^1OC = 576$ ,  ${}^2OC = 6048$ ,  $I_{comp}(\text{func}) = 0.626$ ,  $I_{comp}(\text{loc}) = 0.729$ ,  $I_{vd}(\text{multi}) = 0.766$ .

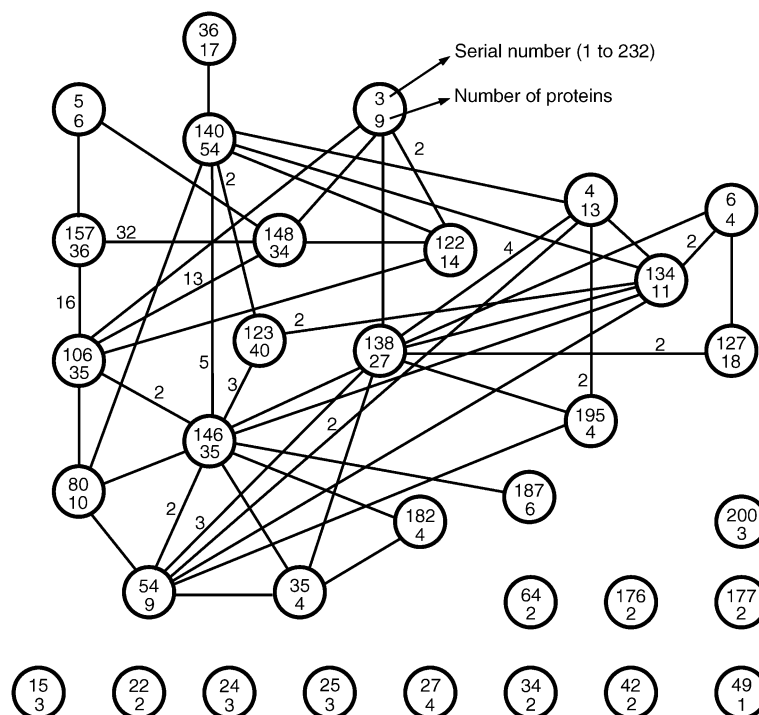
*Fig. 2. Functional organization and complexity descriptors of the yeast proteome.* Some 1,440 proteins are assembled into 232 complexes, which are distributed into nine groups of different biological function (data of *Gavin et al.* [1]). The edge weights of the multigraph indicate that a very large number of proteins can perform more than one biological function.

complexes performing a certain biological function. Two vertices in the graph are connected with an edge, when there is at least one protein that can perform both biological functions. The proteome graph is complete, with a connectedness of 100%, *i.e.*, for any pair of biological function(s), there always exist proteins that can perform both functions. The average intersite distance is equal to one, which indicates that, when the proteome is regarded on a functional level, its ‘small-world network’ reduces to a ‘one-step world’. The edges connecting two functional groups are labeled with the number of proteins they share. One can see that the average number of proteins shared by each of the nine functional groups is about twice (!) the number of proteins in the group itself. Particularly impressive is the large number of proteins shared between the functional groups of ‘RNA metabolism’ and ‘transcription/DNA maintenance/chromatin structure’ (321); between ‘protein synthesis and turnover’ and ‘transcription/DNA maintenance/chromatin structure’ (299); between ‘RNA metabolism’ and ‘protein synthesis and turnover’ (260), *etc.* One may conclude that *the proteome is a highly integrated set of interacting proteins, almost half of the proteins of which can perform more than one biological function.*

The compositional complexity of the yeast proteome, calculated from the distribution of the 232 complexes into nine functional sets of complexes, produces  $I_{\text{comp}}(\text{func}) = 0.626$ . Similarly, one obtains for the distribution of complexes into five subcellular localizations  $I_{\text{comp}}(\text{loc}) = 0.729$ . Both values show a high degree of proteome functional organization. These numbers (as well as all other complexity descriptors) are potentially useful for comparative studies, *e.g.*, between normal and pathophysiological states of the same species for quantitative structure/activity relationship (QSAR) and drug design, or in a normal state of other species for evolutionary analysis, classification purposes, *etc.*

**3.2 The Next Hierarchical Level: Complexity of a Functional or Localization Group of Complexes.** Consider the network characterizing protein synthesis and turnover, which contains 301 proteins organized in 33 complexes (*Fig. 3*). Twelve of the complexes with a total of 29 proteins are idle, whereas the remaining 21 complexes with a total of 124 proteins form a single ‘polycyclic’ component. Every complex in this component shares, on average, 7.6 proteins with other complexes, and can reach every other complex for a little more than two steps. The vertex degree distribution indicates that the complexes numbered 157, 148, and 106 share the largest number of proteins with other complexes (49, 48, and 34, *resp.*). However, one may expect complexes 148 and even 106 to have a stronger effect on protein synthesis and turnover than 157, because they share proteins with six and five other complexes, respectively, while 157 does with only three others. Also a strong effect might be expected from complex 146, which shares 18 proteins with ten other complexes, and from complex 138, which shares 15 proteins with nine other complexes. Thus, one needs both estimates of vertex connectivity – the vertex degree and the multiple vertex degree – to evaluate more adequately the potential importance of each complex for a given biological function.

The central location of the complex in the network is another topological factor to be taken into account, when discussing the role of the complex for a certain biological function. Such a location is favorable, because it determines the complex(es) that can reach all other complexes in the least number of steps. According to the classical graph-center definition [14], eight complexes (146, 138, 140, 106, 80, 3, 122, and 148) should be



Complexity descriptors:  $V = 33$ ,  $E = 125(46)$ ,  $C = 105(26)$ ,  $Conn = 59.5(21.9)\%$ ,  $Cn = 63.6\%$ ,  $\langle a_i \rangle = 11.9(4.38)$ ,  $\langle d_i \rangle = 44.48$ ,  $\langle d \rangle = 2.22$ ,  $W = 467$ ,  $Pl = 2692$ ,  ${}^1OC = 7090$ ,  ${}^2OC = 295996$ ,  $I_{comp} = 0.511$ ,  $I_{vd} = 0.569$ .

Fig. 3. Network and calculated complexity descriptors for the set of protein complexes involved in protein synthesis and turnover (data of Gavin *et al.* [1]). The 33 complexes containing 309 proteins are organized in twelve noninteracting components and a single highly connected component. The values in parentheses are those obtained with the basic vertex/vertex connectivity, whereas those outside the parentheses account for the edge weights as well.

regarded as central because their largest distance to any other complex (called *eccentricity*) is not larger than three. Taking into account the sum of distances to all other vertices (the vertex-distance degree), as a second hierarchical criterion [36–38], complex 146 is identified as the single center of the network, with  $d_{146} = 31$ . Second and third centrally ranked complexes are 138 and 140, with distance degrees of 34 and 36, respectively. The same ranking order of these three complexes is obtained proceeding from the basic vertex degree (the nearest neighbors' number). It is impossible to say *ad hoc* which of these competing factors (the number of shared complexes, the centric location, and the number of complexes accessed in one step), would be more important for a certain biological function. A low total rank in all three orderings seems to be a reasonable practical estimate. For the analyzed protein-synthesis and turnover network, this would classify first complex 146 ( $r = 6$ ), then 138 ( $r = 9$ ), whereas 157 and 106, which have the highest number of shared complexes, would get a considerably higher rank (36 and 14, resp.).

Complexity descriptors calculated for all nine functional groups, as well as for the entire proteome of 232 complexes, are given in the *Table*, and illustrated in *Figs. 4* and *5*.

Most of the complexity descriptors show similar patterns for the nine sets of protein complexes with different biological functions (*Figs. 4* and *5*). All five connectivity descriptors (the total number of shared complexes, connectivity, connectedness,

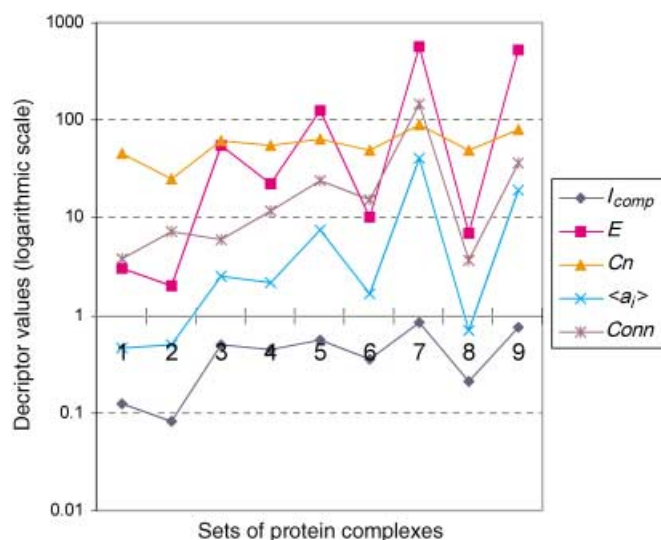


Fig. 4. Similarity in connectivity patterns of nine sets of protein complexes with different biological function. The connectivity parameters used are the total number of edges  $E$ , the average vertex degree  $\langle a_i \rangle$ , the connectivity  $C_n$ , the connectedness  $Conn$ , and the information index on network components  $I_{comp}$ . The highest degree of connectivity is displayed by the set of complexes involved in the RNA metabolism, followed closely by that of the complexes specialized in transcription, DNA maintenance, and chromatin structure.

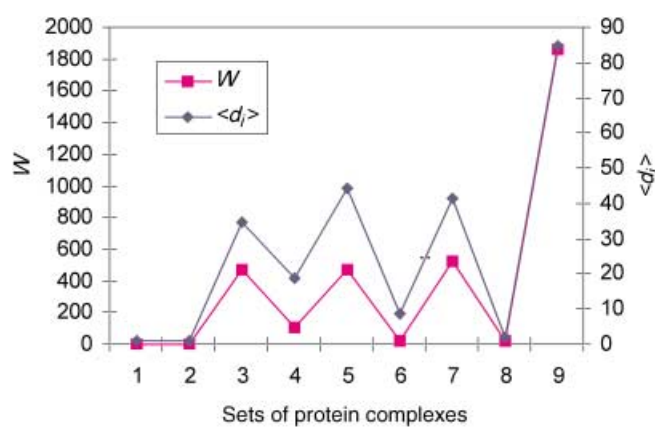


Fig. 5. Similarity in distance patterns of the nine sets of protein complexes with different biological function. The distance parameters used are the total graph distance (the Wiener index  $W$ ) and the average vertex-distance degree  $\langle d_i \rangle$ .

Table. *Quantitative Characteristics of the Functional Groups of Protein Complexes in Yeast Proteome.* All the descriptors have been calculated for the major component of each network.

Functional groups	Complexes	Proteins	Components	$I_{comp}$	$Conn$ ( $Conn'^a$ )	$a_i$ ( $a_i'$ )	$W$	$\langle d \rangle$	$\langle d_i \rangle$	${}^2SC$	${}^1OC$	${}^2OC$	$I_{vd}$	$I_{comp}$
Cell cycle	13	111	$3 \times 2, 7 \times 1$	0.125	– (3.85)	– (0.46)	3	1	1	0	6	0	0	0.498
Cell polarity and structure	8	61	$1 \times 2, 6 \times 1$	0.083	7.14 (3.57)	0.50 (0.25)	1	1	1	1	8	6	0	0.566
Intermediate and energy metabolism	43	221	$21, 3 \times 2, 16 \times 1$	0.421	24.8 (20.5)	4.95 (4.10)	466	2.22	44.4	426	982	13251	0.409	0.376
Membrane biogenesis and traffic	20	147	$11, 9 \times 1$	0.440	40.0 (36.4)	4.00 (3.64)	102	1.85	18.5	99	246	1724	0.427	0.438
Protein synthesis and turnover	33	419	$21, 12 \times 1$	0.554	59.5 (21.9)	11.9 (4.38)	467	2.22	44.5	2692	7090	295996	0.545	0.511
Protein/RNA transport	12	160	$6, 6 \times 1$	0.361	66.7 (40.0)	3.33 (2.00)	26	1.73	8.67	31	96	442	0.447	0.635
RNA Metabolism	28	692	$25, 3 \times 1$	0.862	187.3 (35.3)	45.0 (8.48)	519	1.73	41.5	51034	113296	17473732	0.625	0.565
Signalling	20	125	$2 \times 4, 2, 10 \times 1$	0.208	– (3.68)	– (0.70)	20	1.54	2.00	5	24	25	0.164	0.500
Transcription/DNA repair/ chromatin structure	55	740	$44, 11 \times 1$	0.755	56.9 (24.7)	24.5 (10.6)	1861	1.97	84.6	24036	51338	3800711	0.523	0.453
Entire proteome	232	1440	$199, 2, 31 \times 1$	0.835	19.6(10.8)	38.7 (10.7)	44464	2.26	446.9	374191	778162	136305016	0.480	0.626

a) The  $Conn'$  and  $a_i'$  descriptors are calculated from the nearest-neighbors count, whereas  $Conn$  and  $a_i$  calculation uses edge weights.

average vertex degree, and information for the network components) singled-out the network of complexes controlling the RNA metabolism as having the highest connectivity, followed by those of ‘transcription/DNA maintenance/chromatin structure’, and ‘protein synthesis and turnover’. The same network complexity ordering was produced by the *Platt* index ( $^2SC$ ), and the first- and second-order overall connectivity. In contrast, the networks of ‘cell cycle’, ‘cell polarity and structure’, and ‘signaling’ are very weakly connected (a feature that could change when the entire proteome is analyzed). With the exception of the last three, all other networks have a large connected multi-complex component, and 20–50% single-complex components. The average distance between two complexes in the connected components of all nine networks lies within 1 and 2.2, whereas the maximum distance does not exceed 4, *i.e.*, these are typical ‘small-world’ networks. The small size of the functional group subnetworks does not allow us to verify whether the vertex-degree distribution in them follows the power law, typical for large dynamic networks. The connected component of the RNA metabolism is composed of 25 complexes, each of which shares, on average, *ca.* 40 proteins within the network. Also highly interdependent are the complexes involved in ‘transcription, DNA maintenance, and chromatin structure’, each of which shares on average almost 20 proteins. At the other extreme are the complexes involved in ‘cell cycle’, ‘cell polarity and structure’, and ‘signaling’, each of which shares less than one protein with its counterparts.

**3.3 Proteome Complexity.** Applying the complexity analysis to the proteome as a whole, without accounting for its functional or localization context, one finds that the 232 complexes form one giant component of 199 complexes, another component of two complexes, and 31 single-complex components. The giant component is highly connected. There are 3,853 links between the 199 complexes, each complex linked in average with 21.5 other complexes, sharing with them a total of 38.7 proteins. The average vertex distance in the network is only 2.40, and the maximum number of steps between any pair of complexes (the maximum distance) is 5. The vertex-degree distribution follows the power law: the probability that a randomly chosen vertex from the network has  $k$  nearest neighbors is  $P(k) \sim k^{-\gamma}$  with  $\gamma \approx 1.7$ . The value obtained is within the range of  $1.5 < \gamma < 2.5$  reported for other biological networks [19][39–42] the links in which denote certain type of interaction between the nodes, but not sharing a common element, as is the case in this study. Four complexes, *156*, *163*, *111*, and *161*, share more than 200 proteins each; and another 22 complexes share between 100 and 200 proteins. This indicates that there are many proteins that are shared by more than two complexes, *e.g.*, Kap123 and Lys12, which are shared between 17 and ten complexes, respectively. A total of 39 complexes exhibit the highest centrality, needing only three steps to reach any other complex. Applying our centrality criteria, the three most central complexes are *126*, *163*, and *156*, with total vertex distances of 319, 320, and 322, with 84, 81, and 83 nearest-neighbors’ complexes, and with 188, 218, and 270 shared proteins, respectively.

**Conclusions.** – Different quantitative descriptors of the yeast proteome and its functional and local organization have been discussed. Based on graph and information theory, they assess different complexity components – topological and compositional complexity. Compositional complexity describes the distribution of the network

elements. This could be either a distribution of protein complexes into functional or localization sets of complexes, or a distribution of proteins into complexes or a set of complexes. The information-theoretic measure applied presumes an increasing complexity with the deviation from the most probable even distribution. Topological complexity is regarded as complexity of the graph representing the network. It is essentially based on the graph connectivity, distances, and centrality, as well as on the number of subgraphs. The assumption is that graph complexity increases with vertex/vertex connectivity, which, in turn, results in smaller graph distances. Connectivity is assessed by a number of descriptors, beginning with the number of edges (interactions) between the network elements, the number of cycles, the average vertex degree, and the information index on graph components. Two normalized measures are also used: *connectivity* [3], *Cn*, which is the percentage of vertices having at least one connection with another graph vertex, and *connectedness*, *Conn*, which is the percentage of graph-edges count compared to the maximum possible number of edges at the given number of vertices. Distance descriptors include the sum of distances over the entire graph (the *Wiener* index [16]), the sum of distances of a vertex (*vertex distance degree*), the average graph distance (*average graph radius*), the maximum distance in the graph (*graph diameter*), and the information indices on the distance distributions over vertices and over distance values [32][35]. Distance descriptors are used to determine vertex centrality, making use of a previously established set of criteria [30–32]. Subgraph-based complexity measures [22–24][27–29] proceed from the concept that complexity increases with the number of subgraphs, as well as with the increase in the values of selected graph invariants of each subgraph. Due to the large size of the protein–protein networks, only the first several orders subgraph count [24] and overall connectivity [27][28] are used.

The methodology developed was applied to a complexity assessment of the yeast proteome based on the data published by *Gavin et al.* [1]. We analyzed the proteome functional organization proposed by these authors, beginning with the proteome graph every vertex in which is represented a set of protein complexes with the same biological function. Two vertices are connected by an edge, when the corresponding sets of complexes share at least one protein. The graph thus built is complete (100% connectedness), revealing that, for any pair of biological functions, there always exist proteins that can perform both functions, as well as that numerous proteins can perform more than one biological function.

When a vertex in the proteome graph stands for a protein complex, the resulting large 232-vertex graph can be decomposed into nine smaller subgraphs, representing nine functional sets of protein complexes. The average intersite distance in these graphs varies from 1 to 2.2, and the maximum distance does not exceed 4. For the entire proteome regarded as a graph with 232 vertices, these parameters are 2.4 and 5, respectively. Thus, the proteome and its functional sets of protein complexes belong to the class of ‘small-world’ networks, any pair of vertices of which is connected *via* a small number of steps. Yet, there is an essential difference to mention. Each of these sets of complexes, as well as the proteome itself, is essentially a *disconnected* graph, containing one major component and a number of noninteracting species. The high connectivity in the ‘small-world’ graphs mentioned actually refers to the major component of these disconnected graphs. Since the data available comprises only *ca.* 25% of all proteins in

the yeast proteome [1], it is not yet clear whether the availability of ‘specialized’ protein complexes, which do not share proteins with other complexes, is a general pattern or just an artifact due to the lack of data. Another possible artifact, caused by the insufficient number of proteins, could be the very low connectivity and the lack of a major highly connected component in the functional sets of complexes involved in ‘cell cycle’, ‘cell polarity and structure’, and ‘signaling’. However, from the data available, it is clear that the protein networks controlling the most-vital life functions (like RNA metabolism, transcription, DNA maintenance, and protein synthesis) are based on a high degree of connectivity. This ensures the performance of each specific biological function with a sufficient number of players, even if, for some reason, some of the potential players are off the field.

The question may arise, whether one needs so many different proteome-complexity descriptors, rather than only one or two selected ones. The answer to this question may vary depending on the purpose of the study. If one looks for a quantitative characteristic for classification purposes, then highly discriminating indices like *substructure count* and *overall connectivity* (as well as *walk count* [30][31]) would be sufficient, although the entire set of descriptors could also be used as topological fingerprints. If similarity patterns were the focus of the research, then one might make use of five or six descriptors. For QSAR applications, a reasonable strategy would be to deal with a large variety of models, and then select several of them with the best statistics, make predictions, and dwell on the overlap in the predictions. We regard the descriptors of proteome complexity introduced in this study as a reasonable minimum (particularly, when compared to the existing several hundred topological indices) to be used in proteomics QSAR. Possible applications could include *quantitative* comparative proteome analysis for evolutionary and classification purposes, and, first of all, for assessing the differential expression of proteins in tissues and cells exposed to different physiological conditions, different drugs, or environmental toxins. The quantitative relationships between protein–protein-network structure and the alterations in normal biological activity, as well as the ranking of network vertices according to their centrality, could help in screening potential drug or marker candidates. More-general questions for the protein–protein-network alteration associated with the *Darwinian* evolution, or for the disease-related specific types of proteome network degradation might be addressed along this avenue.

#### REFERENCES

- [1] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. R. Copley, A. Edlmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga, *Nature (London)* **2002**, *415*, 141.
- [2] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreaault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, M. Tyers, *Nature (London)* **2002**, *415*, 180.

- [3] J.-C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schächter, Y. Chemama, A. Labigne, P. Legrain, *Nature (London)* **2001**, 409, 211.
- [4] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, J. M. Rothberg, *Nature (London)* **2000**, 403, 623.
- [5] B. Schwikowski, P. Uetz, S. Fields, *Nat. Biotechnol.* **2000**, 18, 1257.
- [6] C. von Mering, R. Krause, B. Snell, M. Cornell, S. G. Oliver, S. Fields, P. Bork, *Nature (London)* **2002**, 417, 399.
- [7] A. Grigoriev, *Nucleic Acids Res.* **2003**, 31, 4157.
- [8] D. Bonchev, W. A. Seitz, in 'Concepts in Chemistry: A Contemporary Challenge', Ed. Dennis H. Rouvray, Research Studies Press, Taunton, U. K., 1996, 348–376.
- [9] 'Mathematical Chemistry', Vol. VII, 'Complexity in Chemistry', Eds. D. Bonchev, D. H. Rouvray, Taylor and Francis, London, 2003.
- [10] O. N. Temkin, A. V. Zeigarnik, D. Bonchev, 'Chemical Reaction Networks. A Graph Theoretical Approach', CRC Press, Boca Raton, FL, 1996.
- [11] M. Randić, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1330.
- [12] M. Randić, M. Nović, M. Vračko, *J. Proteome Res.* **2002**, 1, 217.
- [13] N. Rashewsky, *Bull. Math. Biophys.* **1955**, 17, 229.
- [14] F. Harary, 'Graph Theory', Addison-Wesley, Reading, MA, 1969.
- [15] C. Shannon, W. Weaver, 'Mathematical Theory of Communications', University of Illinois Press, Urbana, MI, 1949.
- [16] H. Wiener, *J. Am. Chem. Soc.* **1947**, 69, 17.
- [17] 'Topology in Chemistry: Discrete Mathematics of Molecules' (Plenary Lectures at the International Conference Commemorating Harry Wiener, Athens, GA, March 20–24, 2001), Eds. D. H. Rouvray, R. B. King, Horwood, Chichester, 2002.
- [18] D. J. Watts, S. H. Strogatz, *Nature (London)* **1998**, 393, 440.
- [19] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabasi, *Nature (London)* **2000**, 407, 651.
- [20] D. Bonchev, *SAR QSAR Environ. Res.* **2003**, 14, 199.
- [21] D. Bonchev, in 'Handbook of Proteomic Methods', Ed. P. M. Conn, Humana Press, New York, 2003, p. 451–462.
- [22] S. H. Bertz, T. J. Sommer, *Chem. Commun.* **1997**, 2409.
- [23] S. H. Bertz, W. F. Wright, *Graph Theory Notes New York Acad. Sci.* **1998**, 35, 32.
- [24] D. Bonchev, *SAR QSAR Environ. Res.* **1997**, 7, 23.
- [25] J. R. Platt, *J. Chem. Phys.* **1947**, 15, 419.
- [26] S. H. Bertz, *J. Am. Chem. Soc.* **1981**, 103, 3599.
- [27] D. Bonchev, *J. Mol. Graphics Model.* **2001**, 20, 65.
- [28] D. Bonchev, *J. Chem. Inf. Comput. Sci.* **2000**, 40, 934.
- [29] D. Bonchev, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 582.
- [30] G. Rücker, C. Rücker, *J. Chem. Inf. Comput. Sci.* **1993**, 33, 683.
- [31] G. Rücker, C. Rücker, *J. Chem. Inf. Comput. Sci.* **2000**, 40, 99.
- [32] D. Bonchev, 'Information-Theoretic Indices for Characterization of Chemical Structures', Research Studies Press, Chichester, 1983.
- [33] S. Basak, in 'Topological Indices and Related Descriptors in QSAR and QSPR', Eds. J. Devillers, A. T. Balaban, Gordon and Breach, Amsterdam, 1999, p. 563–593.
- [34] A. T. Balaban, T.-S. Balaban, *J. Math. Chem.* **1991**, 8, 383.
- [35] D. Bonchev, N. Trinajstić, *J. Chem. Phys.* **1977**, 67, 4517.
- [36] D. Bonchev, A. T. Balaban, O. Mekenyan, *J. Chem. Inf. Comput. Sci.* **1980**, 20, 106.
- [37] D. Bonchev, O. Mekenyan, A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **1989**, 29, 91.
- [38] D. Bonchev, *THEOCHEM* **1989**, 185, 155.
- [39] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L. Barabasi, *Science* **2002**, 297, 1551.
- [40] A.-L. Barabasi, R. Albert, *Science* **1999**, 286, 509.
- [41] D. Fell, A. Wagner, *Biotechnology* **2000**, 18, 1121.
- [42] A. Wagner, D. Fell, *Proc. R. Soc. London, Ser. B* **2001**, 268, 1803.

Received September 22, 2003