Functional and topological characterization of protein interaction networks

Soon-Hyung Yook[1], **Zoltán N. Oltvai**[2] **and Albert-László Barabási**[1]

[1]Department of Physics, University of Notre Dame, Notre Dame, IN, USA
[2]Department of Pathology, Northwestern University, Chicago, IL, USA

The elucidation of the cell's large-scale organization is a primary challenge for post-genomic biology, and understanding the structure of protein interaction networks offers an important starting point for such studies. We compare four available databases that approximate the protein interaction network of the yeast, *Saccharomyces cerevisiae*, aiming to uncover the network's generic large-scale properties and the impact of the proteins' function and cellular localization on the network topology. We show how each database supports a scale-free, topology with hierarchical modularity, indicating that these features represent a robust and generic property of the protein interactions network. We also find strong correlations between the network's structure and the functional role and subcellular localization of its protein constituents, concluding that most functional and/or localization classes appear as relatively segregated subnetworks of the full protein interaction network. The uncovered systematic differences between the four protein interaction databases reflect their relative coverage for different functional and localization classes and provide a guide for their utility in various bioinformatics studies.

Keywords: Bioinformatics / Protein interaction networks / Scale-free networks

Received	8/5/03
Revised	28/8/03
Accepted	1/9/03

1 Introduction

As protein-protein interactions are central to most biological processes, the systematic identification of all protein interactions is considered a key strategy for uncovering the inner workings of a cell. Consequently, a number of experimental and computational techniques have been developed to systematically determine both the potential and actual protein interactions in selected model organisms, primarily in *Saccharomyces cerevisiae* [1–8]. This proliferation of interest and tools resulted in extensive databases of protein interactions, covering organisms from bacteria to eukaryotes, and fueling research aimed at understanding the large-scale organizing principles of cellular function [9, 10].

As the interactions, in which a given protein participates, are likely to correlate with the protein's functional properties, protein interaction maps are frequently utilized to uncover in a systematic fashion the potential biological role of proteins of unknown functional classification [4, 11]. Also, the topology of the uncovered protein interaction networks may reflect the cell's higher-level functional organization. Yet, despite their clear utility, there is very little understanding to what degree the collected protein network topologies encode such functional information [9]. For example, four different protein interaction maps are currently used, often interchangeably, to approximate the protein interaction network of yeast, but the limitations and quality of the four databases remains poorly studied. Of these, two independently performed systematic two hybrid assays provide us with maps of potential pair wise interactions [1, 2]. In addition, two hand-curated databases, MIPS [12] and DIP [13], collate experimentally determined protein-protein interactions from the literature but use the results of the two-hybrid experiments as well, thus incorporating both demonstrated and potential pairwise protein interactions. While many interactions appear in all four databases, the disparities between the four datasets are notable. For example, the overlap in the interactions identified by the two independent two-hybrid datasets [1, 2] is only between 16–20% [2, 14]. This limited overlap could indicate that the two-hybrid techniques cover

Correspondence: Albert-László Barabási, Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA
E-mail: alb@nd.edu
Fax: +1-574-631-5952

only a small percentage of the potential interactions, or could signal a high rate of false negative and positive interactions [9].

To determine how well the four available databases characterize the protein interaction network of *S. cerevisiae*, here we systematically analyze the relationship between the topology of the obtained protein interaction maps and the known functional properties of the proteins. We start by demonstrating that the four protein interaction networks are characterized by comparable degree and cluster size distributions, indicating that they are described by the same large-scale topology that is best approximated by a hierarchically modular [15] network structure. We also find that despite the potentially small coverage and the ambiguity of the uncovered interactions there are clear correlations between the known functional classification of the proteins and the underlying large-scale topology of protein interaction network in all four datasets. Indeed, most proteins sharing similar functional roles appear segregated on well-defined regions of the protein interaction maps and display a high degree of network based clustering. Similar signatures of network-based segregation are obtained when we study the impact of cellular localization on the network topology, finding that proteins sharing the same subcellular localization form relatively compact subclusters in the protein interaction network. There are noticeable differences, however, in the degree of correlations between function, subcellular localization, and topology characterizing the different databases. The developed methods and subsequent results allow us to uncover the functional relationship between the functional classes and to provide a guide for the utility of the four databases for various bioinformatics studies of the yeast proteome.

2 Methods

The primary focus of the paper is the potential and experimentally determined direct physical interactions occurring between *S. cerevisiae* proteins. The information on protein interactions are deposited in four separate databases: (i) The Database of Interacting Proteins or DIP [13], combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both manually and automatically, using computational approaches that utilize the knowledge about the protein-protein interaction networks. The database downloaded from http://dip.doe-mbi.ucla.edu/ contains information for 5798 yeast proteins (on August 2001) with at least one interaction to other proteins, connected to each other by over 20 000 interactions. (ii) The Munich Information Center for Protein Sequences [12] or MIPS (mips.gsf.de), another hand-curated database for *S. cerevisiae*, (on February 2002), contained information on 6552 proteins, connected *via* 3797 interactions. (iii) The Uetz dataset [1] summarizes the results of the systematic two-hybrid assay, collecting information on 2115 proteins connected *via* 4480 interactions. (depts..washington.edu/sfields/) (iv) The Ito database [2] contains the results of an independent two-hybrid assay. Currently it contains information on 3280 proteins connected *via* 8868 interactions. For completeness we analyzed separately the full dataset, including all detected protein interactions, as well as the higher confidence core data, a subset of the full dataset containing only interactions with more than three interaction sequence tag hits (genome.c.kanazawa-u.ac.jp/Y2H/). Finally, http://us.expasy.org/, http://www.ncbi.nlm.nih.gov/, http://dip.doe-mbi.ucla.edu/ and http://mips.gsf.de/ databases were used to obtain the protein name conventions necessary to compare the different datasets.

3 Results

3.1 Protein interaction databases

For the present study, we focused on four large *S. cerevisiae* protein interaction databases: DIP [13], MIPS [12], which are hand-curated databases, and the two-hybrid dataset collected by Uetz *et al*. [1] and Ito *et al*. [2] The relationship between the four datasets is summarized in Fig. 1 and Table 1, and their particulars are detailed in Section 2. As the two largest datasets, DIP and MIPS, contain interaction data for a significant fraction of the yeast proteins, there is a rather large overlap between them. In contrast, the Uetz and the Ito datasets are subsets of both hand-curated databases. Yet, as noted before [2, 14], the overlap between the Uetz and the Ito data is rather small: less than 30% of the yeast proteins in the Ito data set are found in the Uetz data, and *vice versa*, only about 30% of the Uetz proteins appear in the Ito database as well. At the level of identified protein interactions the differences between the four databases are even more significant (Fig. 1b); for example, only 7% of the interactions identified by the Ito dataset overlap with those present in the Uetz data.

3.2 Large-scale organization of the protein interaction network

Understanding the large-scale organizing principles of protein interaction networks is one of the prominent goals of post-genomic biology. Rapid advances in complex net-

Table 1. Summary of the characteristics of the four studied protein interaction databases

Dataset	Number of proteins in the database (N)	Number of proteins with at least one interaction (N_{int})	Total number of interactions (L)	Largest cluster (LC)		Diameter of the largest cluster	Clustering coefficient (C/C_{rand})	Degree exponent (γ)	Average segregation ($m(1)/m^*$)
				Number of protein in the LC (N_{LC})	Number of interaction in the LC (L_{LC})				
MIPS	6745	2043 (118)	5434	1441	4538	7.71	34.91	2.34	3.28
DIP	5798	5798 (352)	20098	4198	15892	4.9	117.09	2.50	3.48
Uetz	2115	1870 (74)	4480	1458	3941	6.8	54.64	2.32	2.28
Ito	3280	3280 (82)	8868	2840	8371	4.9	36.40	2.44	1.49
Ito core	797	797 (52)	1560	417	1055	6.2	4.94	2.1	7.06

The second column denotes the total number of proteins in the entire dataset (N) while the third column represents the number of proteins which appear in the protein interaction network (N_{int}). In the parenthesis we show the number of proteins which have self-interactions. The fourth column shows the number of interactions between the proteins (L). For the largest cluster (LC) data, N_{LC} denotes the number of proteins in the largest cluster, and L_{LC} represents the number of links in the largest cluster. The diameter denotes the average node-to-node distance for the proteins in the largest cluster being shown in the seventh column. The clustering coefficient is normalized with the clustering coefficient of the random network (see text). The degree exponents are obtained from the relation $P(k) \sim k^{\lambda}$ and the average segregation parameter reflects the tendency of the proteins to be primarily connected to proteins that belang to the same functional class.

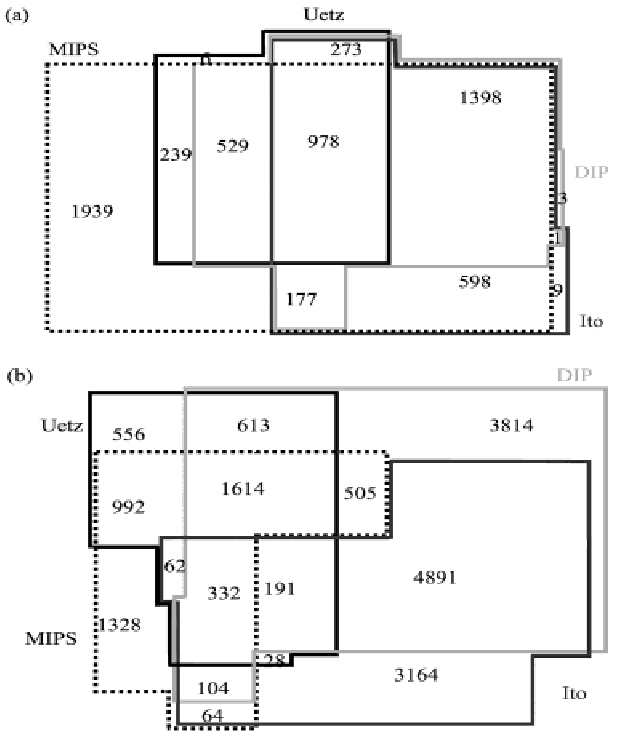


Figure 1. Relationship between the four studied databases (a) at the protein and at (b) the interaction level. The sum of the numbers within each color boundary denotes (a) the total number of proteins or (b) the total number of interactions found in the corresponding database. For example, (a) indicates that the MIPS database has altogether 6745 proteins, of which 1939 proteins with at least one interaction do not show interactions in any other databases. Note that while the overlap between the Uetz and Ito maps, (a) at the protein level is as high as 30%, (b) at the interaction level it is much smaller.

work theory in general have considerably aided this quest [16, 17]. Briefly, for many decades complex networks were modeled at random [18] assuming that a fixed number of nodes (N) are connected by randomly placed links. An important prediction of this model is that the number of nodes with k links follows a Poisson distribution, which implies that most nodes have roughly the same number of links. In contrast, recent studies focusing on large real networks have demonstrated that many of them have a scale-free topology, in which the number of nodes with k links follows a power law distribution, $P(k) \sim k^{-\gamma}$, where γ is the degree exponent [19]. Recent studies have shown the relevance of this type of connectivity for cellular network as well [20–25]. In particular, the protein interaction network generated by the two hybrid approaches has also been found to have a scale-free topology [22, 26]. As scale-free networks are dominated by a few highly connected nodes, or hubs, the inhomogeneous nature of the network topology has important consequences on the robustness and error tolerance of the underlying cellular networks as well [26, 27]. Yet, it is unclear if the scale-free topology is a generic feature of all four protein network maps. To investigate the generality of the scale-free concept in Fig. 2a, we show the degree distribution for all four protein interaction maps on a log-log plot. As the fit indicates, each of the four derived networks have a power law degree distribution, indicating that they are all described by scale-free networks with comparable degree exponent γ (Table 1).

An important question underlying biological organization relates to the potential existence of modules in biological networks. Indeed, following the recent proposal of modu-

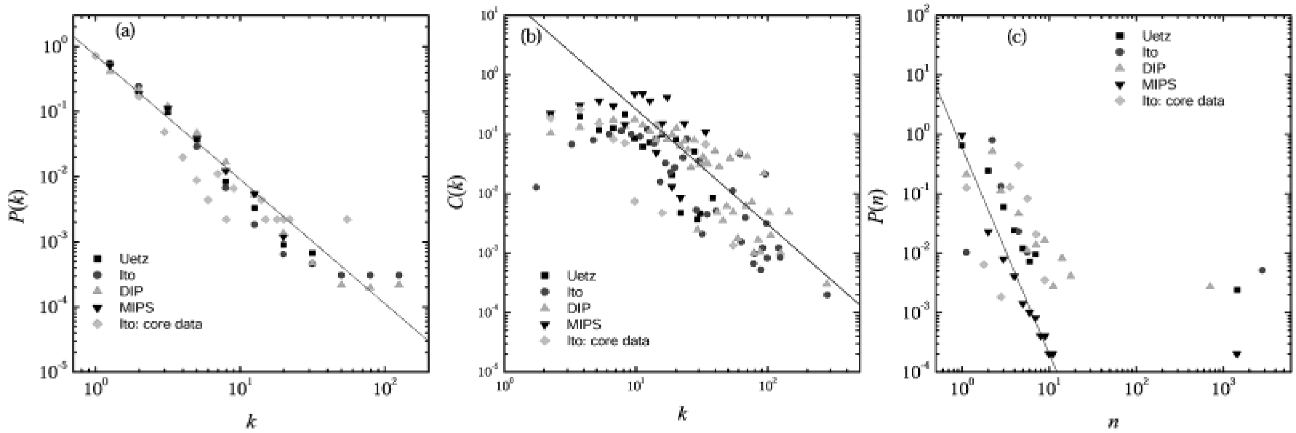


Figure 2. Large-scale characteristics of the protein interaction databases. (a) Degree distribution of the four databases, shown on a log-log plot. Note that all datasets have a power law tail, indicating that the underlying network has a scale-free topology. The solid line is obtained from the fitting to the function $P(k) \sim k^{\gamma}$ to the DIP data, the best fit indicating $\gamma \approx 2.5$ for DIP data set. (b) Distribution of the clustering coefficient for the four studied databases shown on a log-log plot. The straight line has slope -2. (c) Cluster size distribution for the four databases shown on a log-log plot. Apart from the points corresponding to the giant component (for right) the $P(n)$ curves follow a power law. The solid line is obtained from the least square fitting to $P(n) \sim n^{-\alpha}$ for the MIPS dataset, providing $\alpha = 3.4$.

lar biology [28], a series of studies have focused on identifying the biological modules in various cellular networks, ranging from the metabolism [15, 29–31] to genetic networks [2, 32]. Modularity assumes the existence of groups of proteins that work together to achieve some well-defined biological function. For example, it is experimentally well established that protein complexes that act as functional modules carry out many biological functions. From the network perspective these modules should appear as distinct group of nodes that are highly interconnected with each other but have only a few links to nodes outside of the module. Yet, the scale-free topology apparently forbids the existence of independent modules in the network, as the hub proteins' ability to interact with a high fraction of each module's components makes a module's relative isolation all but impossible. Recently, we proposed that the network's scale-free topology can be reconciled with its potential modularity within the framework of a hierarchical modularity [15, 30, 33]. The most important test of such hierarchical modularity is the scaling of the clustering coefficient, C, defined as $C_i = 2n_i/k_i(k_i-1)$ for each node i that has k_i links, where n_i denotes the number of direct links between the k_i neighbors of node i. For the random and the scale-free model the clustering coefficient of a node with k links is independent of k, that is, on average hubs have the same clustering coefficient as small nodes do. In contrast, for a hierarchical network the clustering coefficient $C(k)$ depends on the node's degree as [15, 33–35]

$$C(k) \sim k^{-\beta} \tag{1}$$

where β is the modularity exponent characterizing the network's hierarchical modularity. Therefore, the $C(k)$ function, which can be measured for arbitrary networks [30], can provide direct evidence if the network has a hierarchical modularity. To test the organization of modularity in protein interaction networks we measured the $C(k)$ function for each of the four studied protein network databases. As Fig. 2b shows, we find that $C(k)$ is not independent of k, but can be well approximated by a power law with exponent $\beta \approx 2$, giving direct evidence of hierarchical modularity in protein interaction networks.

Another important property of the currently available protein interaction networks is that they are fragmented into many distinct clusters [11, 22, 26]. Indeed, we find that each of the four databases are dominated by a giant cluster that contains a significant fraction of all connected proteins, such that one can find a path of protein interactions between any two proteins belonging to this giant component. A small fraction of proteins, however, are either completely isolated (*i.e.*, do not have any known interactions to other proteins) or form small islands of isolated groups of interconnected proteins. To characterize the fragmented nature of the protein interaction network we determined the size n of each isolated cluster, and prepared a normalized histogram of the results, obtaining the cluster size distribution. As Fig. 2c shows, each of the datasets have a giant component of approximately 10^3 proteins. However, the giant component coexists with many isolated proteins, somewhat fewer two protein clusters, and even fewer three-protein clusters. If we disregard the giant component, the cluster size distribution follows a power law, $P(n) \sim n^{-\alpha}$ where α is the cluster size exponent, with values ranging between $\alpha \approx 3$–4. This fragmentation could indicate that the existing databases contain only a small fraction of all protein-protein interactions present in *S. cerevisiae*. Indeed, if more protein interactions are uncovered, the giant component is expected to absorb a larger fraction of all proteins, and a fully connected protein network could emerge with a single giant component. Such increase of the giant cluster is a well-known result of random graph theory [18] predicting that as the number of interactions increase in a network with a fixed number of nodes, the isolated clusters will be gradually absorbed by the giant cluster and eventually disappear. Finally, we find that the giant component is typically highly interconnected, resulting in a small node-to-node distance (or diameter). Indeed, the average node-to-node distance for each of the four datasets varies between 4 and 8 (Table 1), indicating that protein interaction networks have small world properties.

In summary, regarding the large-scale topology of protein interaction networks all four databases display the same generic properties: they are all scale-free networks forming a giant cluster accompanied by many small disconnected clusters of proteins; they display a high degree of modularity with a hierarchical organization; and the giant cluster has a small diameter, an indication of its small world property. As these properties are derived from all four databases, they appear to be generic features of the yeast protein interaction network.

3.3 Correlations between topology and functional organization

To correlate the topological and functional properties of the derived protein interaction networks, we utilize the functional classification established by the MIPS database, in which each protein is assigned to one or several of 14 functional classes, based on functional information reported in the literature (Table 2). While a classification scheme into 44 functional classes is also available (www.proteome.com), our choice for the 14-class classification system was motivated by statistical purposes: many functional classes in the 44 class breakdown contain too few proteins to allow us to systematically analyze their segregation and clustering properties.

Table 2. Functional classes based on the MIPS database

ID	Function name	N_i				
		Uetz	MIPS	DIP	Ito	Ito core
0	Metabolism	324	1065	605	541	119
1	Energy	72	252	140	132	31
2	Cell growth, cell division, and DNA synthesis	485	836	586	435	125
3	Transcription	404	793	534	416	140
4	Protein synthesis	88	359	152	160	23
5	Protein destination	278	589	392	320	93
6	Transport facilitation	53	311	139	144	10
7	Cellular transport and transport mechanisms	237	498	313	268	88
8	Cellular biogenesis	80	206	125	99	30
9	Cellular communication/signal transduction	83	135	96	63	20
10	Cell rescue, defense, cell death, and ageing	156	369	231	192	53
11	Ionic homeostasis	30	124	70	61	6
12	Cellular organization	1006	2261	1444	1160	314
13	Classification not yet clear-cut	39	146	65	80	18
14	Unclassified proteins	489	2420	805	1233	267
*	Transposable elements, viral and plasmid proteins	2	116	4	5	0

The table shows the number of proteins in each of the four databases that are known to belong to a given functional class. We neglected the functional class that describes transposable elements, viral and plasmid proteins, as this functional class has less than 10 proteins in the Uetz, Ito and DIP databases, too few for a relevant statistical characterization.

We start from the hypothesis that proteins belonging to the same functional class have a high chance of working together, and thus potentially have a high number of connections between each other. If this were true, we expect the topology of the protein interaction network to be segregated into different functional classes, such that a given protein interacts predominantly with proteins belonging to the same functional class, and only to a lesser degree with proteins belonging to other functional classes. To investigate the validity of this hypothesis for each protein *i* that belongs to functional class λ we define the segregation function, $m_i^\lambda(d)$, as

$$m_i^\lambda(d) = \frac{M_i^\lambda(d)}{M_i(d)} \qquad (2)$$

where $M_i^\lambda(d)$ denotes the number of proteins at distance *d* from protein *i* that belong to the functional class λ and $M_i(d)$ denotes the total number of proteins at distance *d* from protein *i*. As illustrated in Fig. 3a, a protein with links only to nodes in the same functional class has $m_i(1) = 1$ (Fig. 3b), while one that does not have links to any protein of the same functional class has $m_i(1) = 0$. As the topology of the network around a single protein is statistically not representative, it is useful to define $m^{\lambda}(d)$ as the average of $m_i^{\lambda}(d)$ over all proteins *i* that belong to the same functional class λ. Therefore, $m^{\lambda}(d)$ offers a measure of the degree of segregation for the functional class λ. If proteins belonging to a given functional class λ were randomly distributed

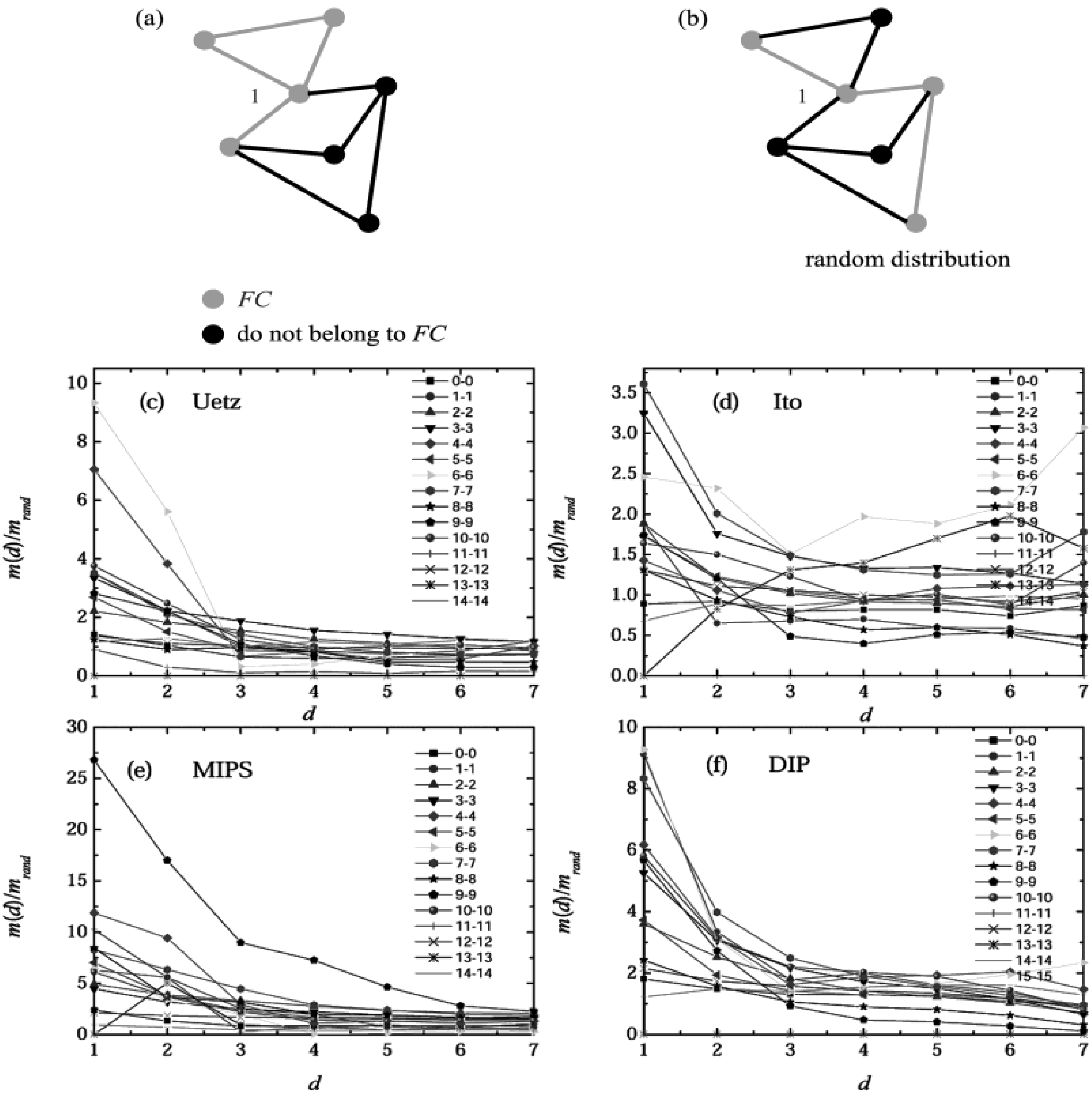


Figure 3. Functional segregation. (a), (b) Schematic illustration of the topological interpretation of the segregation parameter $m(1)$. For example, for protein 1, focusing only on interaction to proteins belonging to the same functional class (shown in light color), we have $k = 4$, $C_1 = 1/3$, $m(1) = 3/4$. Focusing on proteins that are $d = 2$ distance from protein 1, we have $m(2) = 0$. In (b) we reorganized randomly the same number of proteins and links. In this random configuration we find $C_1 = 0$; $m(1) = 2/4 = 1/2$; $m(2) = 1/2$. (c)–(f) The relative segregation functions for the four databases. Each curve corresponds to a different functional class, the numbers and the corresponding functional classes being listed in Table 2. The four panels describe the different dataset, *i.e.*, (c) Uetz, (d) Ito complete, (e) MIPS, and (f) DIP.

in the network, then $m^\lambda(d)$ should be independent of the distance d, equal to $m^\lambda{}_{\text{rand}}$, where $m^\lambda{}_{\text{rand}}$ is the average density of proteins that belong to the functional class λ, given by $m^\lambda{}_{\text{rand}} = N^\lambda/N$, where N^λ denotes the total number of proteins that belong to the functional class λ, and N is the total number of proteins in the protein network. In contrast, if proteins belonging to the functional class λ have a tendency to cluster together, we expect the associated $m^\lambda(d)$ function to monotonically decrease, converging for large d to $m^\lambda{}_{\text{rand}}$.

The $m^\lambda(d)$ curves obtained for each of the 14 functional classes are shown in Figs. 3c–f for the four protein interaction networks. As the number of proteins differ between the functional classes, one expects large, functionally irrelevant variations in $m^\lambda(d)$. To offset these variations, in Figs. 3c–f we plot the relative segregation function $m^\lambda(d)/m^\lambda{}_{\text{rand}}$ for all four datasets. The ratio $m^\lambda(d)/m^\lambda{}_{\text{rand}} > 1$ if the proteins belonging to λ display measurable topological segregation. For most functional classes we observe that $m^\lambda(d)/m^\lambda{}_{\text{rand}} \gg 1$ for small d, and decreases rapidly with d, reaching the asymptotic limit $m^\lambda(d)m^\lambda{}_{\text{rand}} \sim 1$ for $d \geq 3 \sim 4$. This indicates that most functional classes display some degree of topological localization within the protein interaction network, *i.e.*, the immediate neighbors of a given protein belong with high probability to the same functional class. For some functional classes the segregation function for small d is over 10, implying that the proteins belonging to this class are 10 times more likely to have neighbors that belong to the same functional class than proteins randomly placed in the network. For example, this high degree of segregation is seen for proteins contributing to transport facilitation (#6 in Table 2) in the Uetz data, or cellular communication and signal transduction (#9) in the MIPS database.

To compare directly the four protein interaction networks in Fig. 4a we plot $m(1)/m_{\text{rand}}$ for each of the four datasets and the 14 functional classes. The results obtained for the

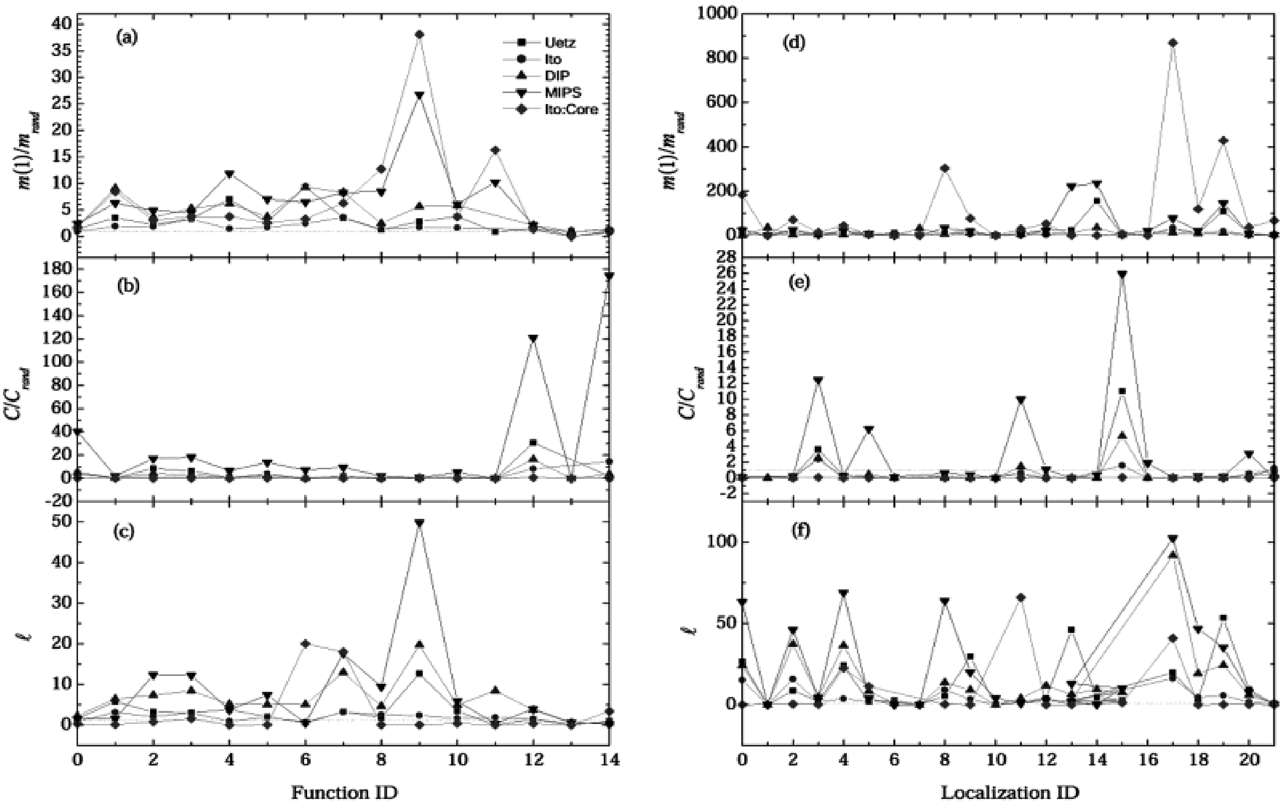


Figure 4. Characterization of the segregation properties of proteins classified based on their functional class (a–c) or subcellular localization (d–f). In (a–c) on the horizontal axis we show the number corresponding to the various functional classes described in Table 2. (a) $m(1)/m(1)_{\text{rand}}$ ratio for each functional class, shown separately for each of the five studied databases. (b) Relative clustering coefficient C/C_{rand} for each functional class, shown separately for all five databases. (c) Relative number of links between proteins belonging to the same functional class, L/L_{rand}, for each of the 14 functional classes, shown separately for the five databases. (d–f) The same quantities as in (a-c) but characterizing proteins sharing the same subcellular localization. The horizontal axis, therefore denotes the cellular localization classes listed in Table 3. The color code is the same as in (a), the plots showing separately the data for each localization class (horizontal axis) and each database.

four datasets correlate with each other: a high degree of functional segregation of one dataset is typically reflected as some degree of segregation in the other datasets as well. We observe a high degree of segregation for the Uetz, MIPS, DIP, and the core Ito data. In contrast, with a few exceptions, the extended Ito dataset displays a smaller degree of functional segregation, while the DIP and the core Ito datasets have the highest $m(1)/m_{\mathrm{rand}}$ coefficient for most functional classes.

If nodes belonging to a given functional class form cohesive groups within the protein interaction network, they should display a high degree of clustering. The degree of clustering of a complex network is often characterized by the clustering coefficient, *C*, discussed above. To determine the degree of clustering for each functional class we restricted the network to the nodes belonging to a given functional class and direct links between them, and measured the average clustering coefficient for the obtained functional sub-graph. As often this sub-graph is rather fragmented (particularly for functional classes with smaller number of nodes), the value of the clustering coefficient by itself is not particularly revealing. To obtain a meaningful measure, we calculate the relative clustering coefficient, $c^{\lambda} = C^{\lambda}/C^{\lambda}_{\mathrm{rand}}$ for each functional class (Fig. 4b), where to determine $C^{\lambda}_{\mathrm{rand}}$ we randomly distribute on the network N^{λ} proteins (*i.e.*, assign randomly chosen proteins to the functional class λ, without altering the network topology), and measure $C^{\lambda}_{\mathrm{rand}}$ for the obtained random subnetwork λ. One can notice the high degree of correlation between the results obtained for all datasets. The pattern seen in Fig. 4a is evident here as well, the degree of clustering observed for the MIPS, DIP and Uetz datasets being very high. Overall the MIPS database has the highest relative clustering coefficient for most functional classes.

Finally, another measure of a network's functional segregation can be obtained by determining the number of direct links between proteins that belong to the same functional class. Let us consider an arbitrary functional class λ, and denote by L^{λ} the number of direct links between proteins that belong to λ. To obtain a meaningful measure of the topological cohesiveness of functional class λ, we calculate the ratio $\ell \equiv L^{\lambda}/L^{\lambda}_{\mathrm{rand}}$, where $L^{\lambda}_{\mathrm{rand}}$ is the number of direct links between proteins of functional class λ if the proteins of λ are placed randomly on the network, without altering the network's topology. A ratio $\ell^{\lambda} = 1$ implies that the proteins belonging to λ are randomly distributed in the network. A ratio of 10, however, indicates that there are 10 times more internal links within the functional class λ than expected for a random protein distribution. The results again indicate large deviations from a random distribution for the DIP, MIPS, Uetz, and core Ito datasets, and weak segregation for the complete to data.

The combination of the results of Fig. 3 offer a rather detailed characterization of each of the four protein interaction networks and allow us to uncover systematic differences between the different functional classes. For example, we found that proteins responsible for cellular communication and signal transduction (#9, Table 2) show a very high segregation parameter in the MIPS database, indicating that the neighbors of a protein contributing to cellular communication are 26 times more likely to belong to the same functional class, and they interact only with such proteins. This finding is corroborated by Fig. 4c as well. The clustering coefficient of this class is not remarkable, however (Fig. 4b). Therefore, the proteins belonging to this functional class mostly interact with each other but they form a loose collection of nodes, with a small degree of clustering. In contrast, the functional class responsible for cellular organization (#12, Table 2) has a very high clustering coefficient in all databases, the corresponding proteins forming strongly interconnected clusters. It does not have an unusually high segregation parameter, however, indicating that in addition to the direct links within the same functional class, cellular organization proteins also interact with a large number of proteins from other functional classes.

3.4 Correlation between topology and cellular localization

Depending on the functional role they play, proteins are often localized in spatially distinct areas of the cell. This spatial compartmentalization is particularly prominent for eukaryotes, and is expected to leave its mark on the topology of the protein interaction network as well: a protein localized in the nucleus is more likely to interact with another nuclear protein than with those localized at the cell wall. To investigate the correlation between cellular localization and the protein network topology, we assigned each protein its cellular location based on a 28 subcellular localization classes (Table 3) obtained from the Proteome database (www.proteome.com).

To characterize the correlations between the topology of protein interaction networks and the known subcellular localization properties of the yeast proteome, we used the quantities developed earlier, measuring for each localization class λ the function $m^{\lambda}(1)/m^{\lambda}_{\mathrm{rand}}$, $C^{\lambda}/C^{\lambda}_{\mathrm{rand}}$, and $L^{\lambda}/L^{\lambda}_{\mathrm{rand}}$. As Fig. 4 demonstrates, our measurements indicate that most localization classes appear segregated in the protein interaction network. In particular, proteins belonging to a few classes, such as those localized in the mitochondrial matrix or outer membrane, or nuclear pore, show an over 100 time increase in their localization coefficient compared to the randomly distributed reference set. In addition, we observe correlations between the degree of

Table 3. Classification of the *S. cerevisiae* proteins in cellular localization classes, based on the Proteome (www.proteome.com) database

ID	Localization	N_i				
		Uetz	Ito	MIPS	DIP	Ito core
0	Bud neck	37	28	53	39	11
1	Cell wall	13	39	68	34	3
2	Centrosome/spindle pole body	54	44	70	49	19
3	Cytoplasmic	298	385	747	425	102
4	Cytoskeletal	83	50	100	79	18
5	Endoplasmic reticulum	101	119	225	136	25
6	Endosome/endosomal vesicles	20	17	36	22	4
7	Extracellular (excluding cell wall)	5	13	24	11	0
8	Golgi	50	43	93	48	18
9	Lysosome/vacuole	30	49	90	54	15
10	Microsomal fraction	7	13	19	12	2
11	Mitochondrial	95	225	442	238	32
12	Mitochondrial inner membrane	30	72	146	76	7
13	Mitochondrial matrix	14	37	68	39	6
14	Mitochondrial outer membrane	13	16	30	17	1
15	Nuclear	590	598	1123	781	188
16	Nuclear nucleolus	36	67	132	60	8
17	Nuclear pore	36	31	54	46	22
18	Other vesicles of the secretory/endocytic pathways	34	37	65	41	13
19	Peroxisome	16	30	49	31	11
20	Plasma membrane	70	115	234	134	22
21	Unspecified membrane	62	123	275	103	19
*	Cell ends	3	5	6	5	1
*	Contractile ring	1	2	2	2	0
*	Lipid particles	0	10	13	7	3
*	Mitochondrial intermembrane space	3	7	13	7	1
*	Nuclear matrix	4	3	8	5	2
*	Nuclear transport factor	1	1	1	1	0
*	Secretory vesicles	9	5	11	9	2

The table gives the number of proteins belonging to each of the cellular localization classes in each of the four protein databases. We eliminated from our analysis seven localization classes, each containing ten or less proteins in at least three databases, as they did not provide enough data points for reliable statistical study.

localization observed in the four studied datasets. In this case, however, the MIPS database stands out, as it shows a higher degree of segregation than any of the other databases. Most importantly, the fact that the segregation and clustering parameters are significantly higher than one for most functional classes indicates that the topology of the protein interaction network reflects, to a considerable degree, the cell's physical compartmentalization.

Some interesting cases are observed in this case as well. Proteins localized in the mitochondrial outer membrane (#14, Table 3) display a very high degree of segregation ($m(1)/m^* \approx 250, 150, 40$ for MIPS, Uetz, Ito, respectively), yet the L/L_{rand} parameter is not particularly high and the clustering coefficient of this class is not remarkable either. Therefore, while the proteins of this class interact predominantly with each other, they do not form a highly interconnected cluster.

3.5 Relationship between functional and localization classes

The segregation of the various functional classes in separate regions of the protein interaction network inspires a new question: how do these functional classes relate to each other? That is, knowing the overall topology of the protein interaction network, can we establish the relationship between the different cellular functions? If the proteins were to interact only with proteins belonging to the same functional class, the protein interaction network should be broken into islands corresponding to the different functional classes. This is not the case, however, as there are a considerable number of interactions between proteins belonging to different functional classes [11]. The number of links between proteins of two functional classes offers a measure to what degree proteins from two functional classes may act together within functional modules. Thus our goal is to use this measure to derive a global map of potential functional relationships within the yeast proteome.

To determine the degree to which proteins of functional class λ are related to proteins of class ϕ we measure the $\ell(\lambda, \phi)$ coefficient, defined as

$$\ell(\lambda, \phi) = \frac{L^{\lambda,\phi} + L^{\phi,\lambda}}{L^{\lambda} + L^{\phi}} \qquad (3)$$

where $L^{\lambda,\phi}$ is the total number of links that proteins of class λ have to protein members belonging to functional class ϕ and where L^{λ} (L^{ϕ}) is the total number of links between the proteins of functional class λ (ϕ). The ($L^{\lambda,\phi}/L^{\lambda}$)/($L^{\lambda,\phi}_{\text{rand}}/L^{\lambda}_{\text{rand}}$) matrices obtained for the four protein interaction networks are shown in Fig. 5. They indicate that the

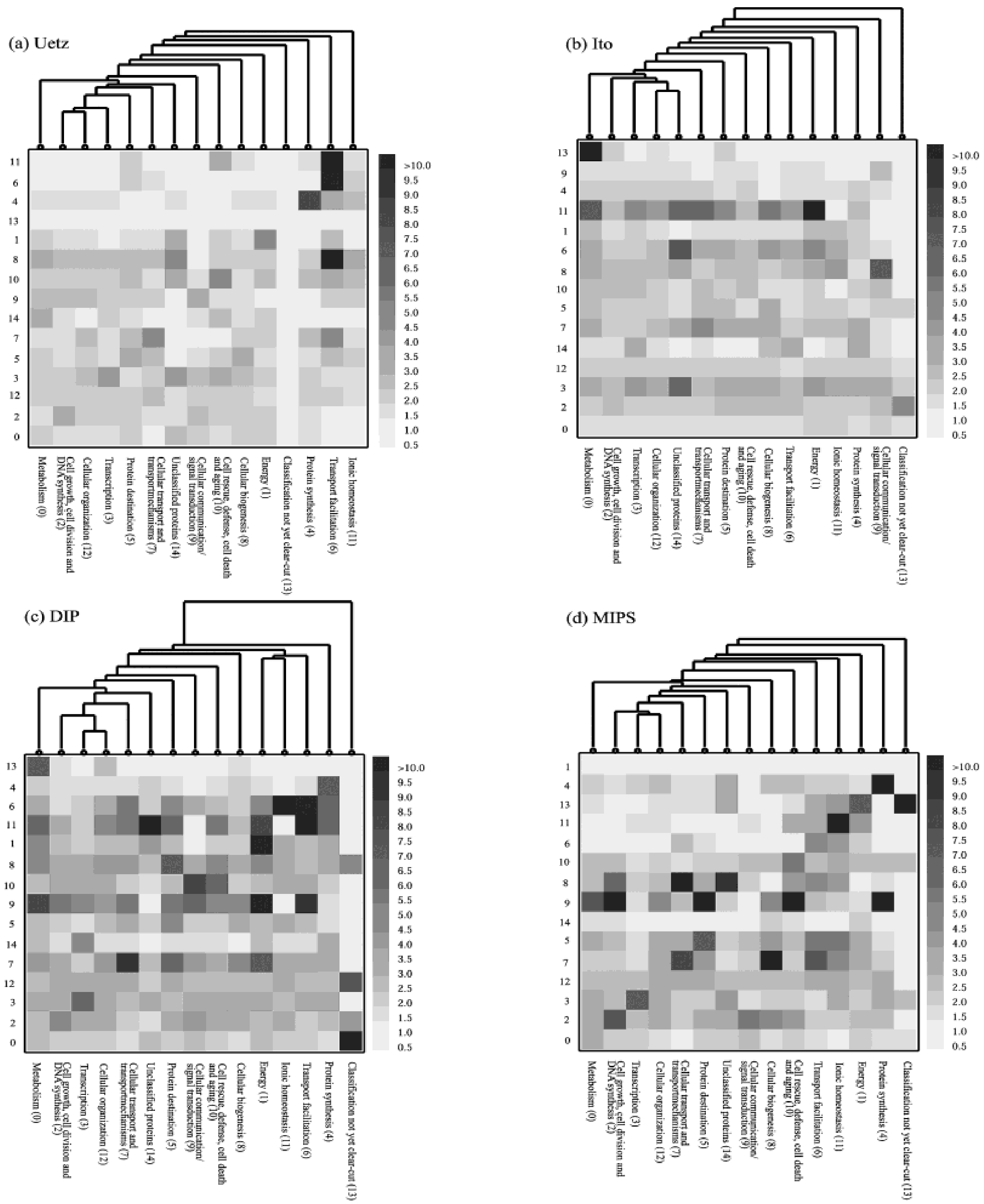


Figure 5. $(L^{\lambda,\phi}/L^{\lambda})/(L^{\lambda,\phi}_{\text{rand}}/L^{\lambda}_{\text{rand}})$ matrix, representing the relationship between the different functional classes based on (a) Uetz, (b) Ito, (c) MIPS, and (d) DIP. The higher the $(L^{\lambda,\phi}/L^{\lambda})/(L^{\lambda,\phi}_{\text{rand}}/L^{\lambda}_{\text{rand}})$ coefficient, the more interactions are detected between the two functional classes. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different functional classes.

relationships between the different classes vary widely: we observe strong ties between some functional classes, while others appear only weakly related.

To uncover the relationship between these functional classes, we applied a minimum linkage clustering algorithm [36] using the quantity $1/\ell(\lambda,\phi)$ as the distance metric between classes λ and ϕ. The algorithm places close to each other the functional classes that are topologically closely related. A hierarchical tree, generated by the clustering process, summarizes the relationship between the different functional classes. At a first glance it is evident that the hierarchical trees obtained for the four databases agree on some generic features of the cell's internal organization. Indeed, all databases indicate that protein function for the classes #12 (cellular organization) and #3 (transcription) belong to the more connected core of the network, closely traced by proteins belonging to the classes #2 (cell growth, cell division and DNA synthesis) #7 (cellular transport and transport mechanisms), and #5 (protein destination). The rest of the functional classes surround this core in an onion-like fashion. All four databases agree regarding the two classes that show the smallest degree of interaction with any other class (and thus are delegated to the outer branches of the hierarchical tree): these are proteins whose classification is not yet clear-cut (#13) and protein synthesis (#4).

We can perform a similar clustering based on the overlap between the proteins belonging to different cellular localization. For this, we measure again the ℓ (λ,ϕ) parameters defined above, but here λ and ϕ denote different localization classes (see Table 3). The results, summarized in Fig. 6, indicate again a relative agreement between the relationships predicted by the four databases. First, the protein interaction network appears to be organized around nuclear proteins (#15), which interact closely with mictochondrial outer membrane proteins (#14) in the Uetz, Ito, and MIPS databases, and rather closely in the DIP data as well. The two other localization classes that are always found in the vicinity of these two core classes include cytoplasmic (#3) and nuclear pore proteins (#17). The hierarchical trees are also consistent regarding the protein groups that are far from the core: extracellular or microsomal fraction proteins (#7 and 10) are clustered together and are far from the rest of the functional classes in most datasets.

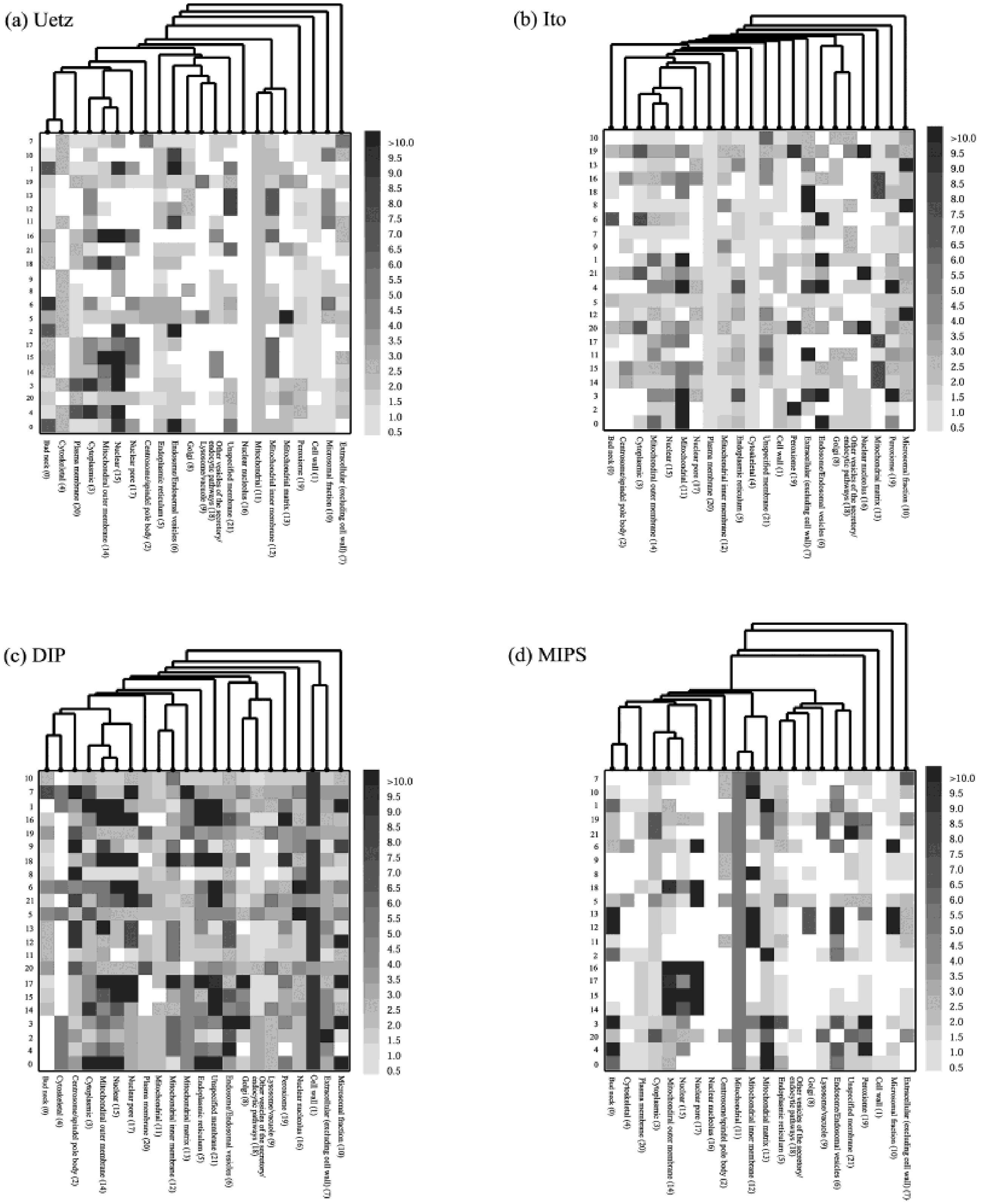
4 Discussion

Uncovering the large-scale properties of protein interaction networks potentially offers an increased understanding of the system level properties of living organisms. Two questions are of primary importance from this perspective: (i) understanding the network's large-scale organization, and (ii) understanding how do these large-scale properties reflect the functional properties of the cellular compartments. The increasingly extensive protein interaction databases, together with the functional annotation of the different proteins, allow us to address these questions in a systematic manner. In the following, we briefly summarize our findings and discuss their implications on our ability to use these databases for various bioinformatics purposes.

4.1 Large-scale organization

Our results offer convincing evidence that networks deduced from the four protein interaction databases have the same large-scale topology. Indeed, each database generates a scale-free network, with embedded hierarchical modularity. We find that the scaling exponents characterizing both the degree distribution $P(k)$ and the modularity distribution $C(k)$ are comparable. As each of the four databases is incomplete, we need to ask if a more complete dataset would change these conclusions. The extensive studies on scale-free networks indicate that this is unlikely [16, 17]: if the underlying network is scale-free, a restricted network, obtained by randomly sampling the links of the scale-free network, will also stay scale-free. In contrast, it is impossible to obtain a scale-free network from the incomplete but random sampling of a network that does not have a power law degree distribution.

Several investigators have proposed that the observed scale-free nature of the protein interaction map is the result of gene duplication, a process frequently occurring during evolution [37–40]. Each gene duplication event leads to a new protein that interacts with the same proteins as the protein product of the original duplicated gene. Proteins that have a large number of links to other proteins are more likely to be connected to a duplicating gene, therefore, they will be more likely to gain new interactions to the newly created protein. This subtle effect leads to both growth (after each gene duplication the network has an additional node, thus the network expands) and preferential attachment (highly connected proteins increase their number of interactions faster than their less connected counterparts, as they are more likely connected to a randomly duplicating protein), the two necessary ingredients for the appearance of a scale-free network [19, 41]. While the microscopic parameters required to predict the precise value of the scaling exponents are still unknown, gene duplication does offer the conceptual framework to understand the origin of the scale-free behavior observed in protein interaction networks. The presence of the power



(a) Uetz

(b) Ito

(c) DIP

(d) MIPS

Figure 6. $(L^{\lambda,\phi}/L^{\lambda})/(L^{\lambda,\phi}_{\text{rand}}/L^{\lambda}_{\text{rand}})$ matrix, representing the relationship between the different subcellular localization classes based on (a) Uetz, (b) Ito, (c) MIPS, and (d) DIP. The higher the $(L^{\lambda,\phi}/L^{\lambda})/(L^{\lambda,\phi}_{\text{rand}}/L^{\lambda}_{\text{rand}})$ coefficient, the more interactions are detected between the two localization classes. The top of each figure shows the tree generated by the hierarchical clustering process, quantifying the overall relationships between different localization classes.

law degree distribution in all four databases (Fig. 2a) supports the expectation that the scale-free topology is a generic feature of the protein interaction network.

The fragmentation of the network into separate, isolated clusters, however, are much more sensitive to potential data incompleteness. Recent models addressing the potential origin of the scale-free topology in protein interaction networks indicate that the observed fragmentation could be an intrinsic property of the evolutionary processes leading to the protein interaction networks [42]. Indeed, the divergence of the duplicated protein sequences by mutations could lead to the loss of interactions between a protein and its interaction partners. If an isolated protein is duplicated, several subsequent duplication events could lead to the emergence of an isolated cluster of proteins. The analytical results indicate that the network emerging as a result of gene duplication and loss of interactions due to mutations can develop a power law cluster size distribution, whose exponent depends on the rate at which proteins add links to other proteins during evolution [42]. Thus, the power law cluster size distribution seen in all four databases could be another consequence of gene duplication and divergence [42]. In the absence of a precise knowledge of gene duplication and divergence rates it is impossible to predict whether the final network should be fully connected or fragmented.

4.2 Function and cellular localization

The three quantities we introduced to quantify the relationship between topology and function/localization allowed us to compare the segregation properties of the four protein interaction networks. The results indicate that the four databases show different strengths in different functional or localization classes. Despite these differences, the two hand-curated databases, DIP and MIPS, display a higher degree of correlation between the network structure and functional/localization based classification than the two two-hybrid datasets. Overall, the MIPS dataset shows the highest degree of functional localization in most functional classes but the DIP dataset often offers a higher degree of cellular localization-based segregation. The weakest correlation between topology and functional and localization-based classification are observed in the complete Ito dataset, but the core Ito data displays correlation comparable to that observed in the Uetz dataset. These results indicate that hand-curated databases not only offer a higher number of interactions but the structure of the protein interaction network reflects better the functional and localization features of the proteins. Therefore, these databases offer a better starting point for bioinformatics studies.

Naturally, the studied databases represent our current knowledge about protein interactions. While *S. cerevisiae* represents one of the most studied organism, the differences between the functional and localization based characteristics of the four studied databases offer a glimpse how incomplete these databases are. This incompleteness comes from two sources: the absence of many potential interactions, and the presence of false positives. While new research continues to add new interactions to these databases, thus gradually addressing the issue of data incompleteness, the presence of false positives will be much harder to eliminate. In addition, the biological literature and the curating efforts tend to focus on scientifically and commercially more interesting subsets of proteins, such as, *e.g.*, signaling pathways. Therefore, certain functional subclasses are better mapped than others and even within a given class some proteins are better characterized than others. It is also a challenge to integrate into these databases the results of the interactions generated by mass spectroscopic studies on protein complexes [43, 44]. Indeed, two hybrid measurements offer information only on pairwise interactions. If two proteins cannot bind together without the help of a third protein, two hybrid datasets will likely not indicate a potential interaction between them. Recently, two groups have provided information on the composition of hundreds of protein complexes under a given growth condition [43, 44]. Each of the components of a given complex is therefore a potential interaction partner [45]. Including the complex information as pairwise interactions in the databases, however, is misleading, as being part of the same complex is not sufficient to establish a pairwise interaction. Yet, the emergence of these new datasets could help to strengthen the validity of known interactions, and offer the hope that with time we will acquire a quite complete map of protein interactions in such simple organism as yeast, serving as a starting point for a better understanding of its functional architecture.

This work was supported by grants from the National Institute of Health and the Department of Energy to A.-L. B & Z.N.O. and National Science Foundation to A.-L. B. Correspondence and requests for materials should be addressed to A.-L. B. (alb@nd.edu) or Z.N.O. (zno008@northwestern.edu)

5 References

[1] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. *et al.*, *Nature* 2000, *403*, 623–627.

[2] Ito, T., Muta, K. S., Ozawa, R., Chiba, T. *et al.*, *PNAS* 2000, *97*, 1143–1147.

[3] Dress, B. L., Sundin, B., Brazeau, E., Caviston, J. P. *et al.*, *J. Cell Biol.* 2001, *154*, 549–576.

[4] Tong, A. H. Y., Drees, B., Nardelli, G., Bader, G. D. *et al.*, *Science* 2002, *295*, 321–324.

[5] Matthews, L. R., Vaglio, P., Reboul, J., Ge, H. *et al.*, *Genome Res.* 2001, *11*, 2120–2126.

[6] Bock, J. R., Gough, D. A., *Bioinformatics* 2001, *17*, 455–460.

[7] Aloy, P., Russell, R. B., *PNAS* 2002, *99*, 5896–5910.

[8] Gomez, S. M., Rzhetsky, A., *Pacific Symposium on Biocomputing* 2002, *7*, 413–424.

[9] von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.*, *Nature* 2002, *417*, 399–403.

[10] Koonin, E. V., Wolf, Y. I., Karev, G. P., *Nature* 2002, *420*, 218–223.

[11] Schwikowski, B., Uetz, P., Fields, S., *Nature Biotechnol.* 2000, *18*, 1257–1261.

[12] Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G. *et al.*, *Nucleic Acids Res.* 2002, *30*, 31–34.

[13] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K. *et al.*, *Nucleic Acids Res.* 2000, *28*, 289–291.

[14] Hazbun, T. R., Fields, S., *PNAS* 2001, *98*, 4277–4278.

[15] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. *et al.*, *Science* 2002, *297*, 1551–1555.

[16] Albert, R., Barabási, A.-L., *Rev. Mod. Phys.* 2002, *74*, 47–97.

[17] Mendes, J. F. F., Dorogotsev, S. N., *Adv. Physics.* 2002, *51*, 1079–1187.

[18] Bollobás, B., *Random Graphs*, Academic Press, London 1985.

[19] Barabási, A.-L., Albert, R., *Science* 1999, *286*, 509–512.

[20] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. *et al.*, *Nature* 2000, *407*, 651–654.

[21] Wuchty, S., *Mol. Biol. Evol.* 2001, *18*, 1694–1702.

[22] Wagner, A., *Mol. Biol. Evol.* 2001, *18*, 1283–1292.

[23] Wagner, A., Fell, D., *Proc. Roy. Soc. London Series B* 2001, *268*, 1803–1810.

[24] Park, J., Lappe, M., Teichmann, S. A., *J. Mol. Biol.* 2001, *307*, 929–938.

[25] Featherstone, D. E., Broadie, K., *Bioessays* 2002, *24*, 267–274.

[26] Jeong, H., Mason, S. P., Barabási, A.-L., Oltvai, Z. N., *Nature* 2001, *411*, 41–42.

[27] Albert, R., Jeong, H., Barabási, A.-L., *Nature* 2000, *406*, 378–382.

[28] Hartwell, L. H., Hopfield, J. J., Leibler, S., Murray, A. W., *Nature* 1999 *402*, C47–C52.

[29] Holme, P., Huss, M., Jeong, H., *Bioinformatics* 2003, *19*, 532–538.

[30] Ravasz, E., Barabási, A.-L., *Phys. Rev. E* 2003, *67*, 026112-1–026112-7.

[31] Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. *et al.*, *Bioinformatics* 2002, *18,* 351–361.

[32] Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O. *et al.*, *Nat. Genet.* 2002, *31*, 370–377.

[33] Barabási, A.-L., Ravasz, E., Vicsek, T., *Physica A* 2001, *299*, 559–564.

[34] Dorogovtsev, S. N., Goltsev, A. V., Mendes, J. F. F., *Phys. Rev. E* 2002, *65*, 066122-1–066122-4.

[35] Szabo, G., Alava, M., Kertesz, J., *Phys. Rev. E* 2002, *67*: 056102-1– 056102–5.

[36] Jackson, B. B., *Multivariate Data Analysis*, Richard D. Irwin, Homewood, IL, USA 1983.

[37] Sole, R. V., Pastor-Satorras, R., Smith, E. D., Kepler, T., *Adv. Complex Syst.* 2003, in press.

[38] Qian, J., Luscombe, N. M., Gerstein, M., *J. Mol. Biol.* 2001, *313*, 673–681.

[39] Vazquez, A., Flammini, A., Maritan, A., Vespignani, A., *ComPlexUs* 2003, *1*, 38–44.

[40] Chung, F., Lu, L., Dewey, T. G., Galas, D. J., *J. Comp. Biol.* 2003, *10*, 677–687.

[41] Barabási, A.-L., Albert, R., Jeong, H., *Physica A* 1999, *272*, 173–187.

[42] Kim, J., Krapivsky, P. L., Kahng, B., Redner, S., *Phys. Rev. E* 2002, *66*, 055101-1–055101-4.

[43] Gavin, A. C., Bösche, M., Krause, R., Grandi, P. *et al.*, *Nature* 2002, *415*, 141–147.

[44] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D. *et al.*, *Nature* 2002, *415*, 180–183.

[45] Dezsö, Z., Oltvai, Z. N., Barabási, A. L., *Gen. Res.* 2003, *13*, 2450–2454.

6 Addendum

The Ito core data has 797 proteins and 1560 interactions.

The functional segregation $m^{\lambda}(d)/m^{\lambda}_{rand})$ for Ito core data, corresponding to Fig. 2 in the text, is shown below.

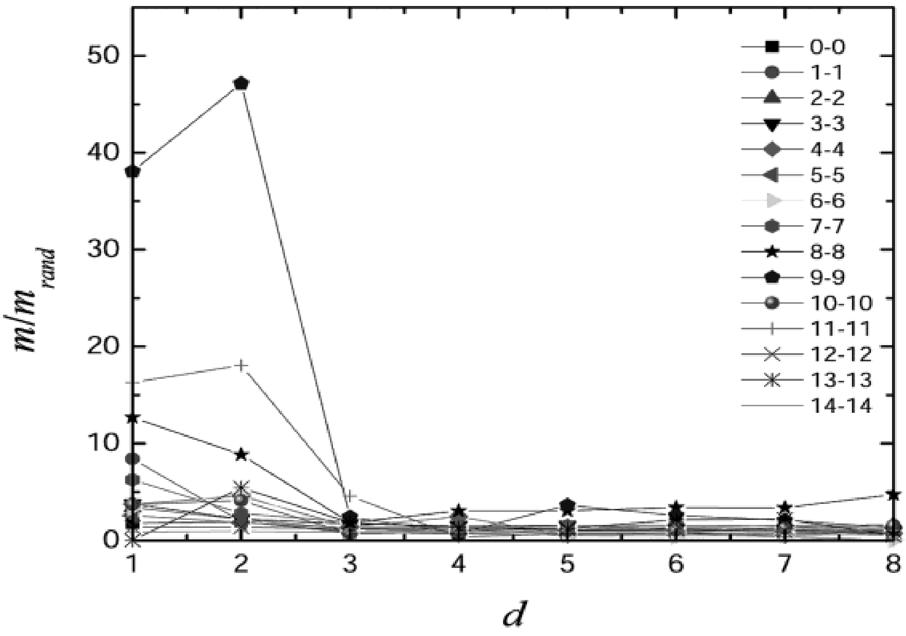


Figure 1. Functional segregation of Ito core data. Each curve corresponds to a different functional class.

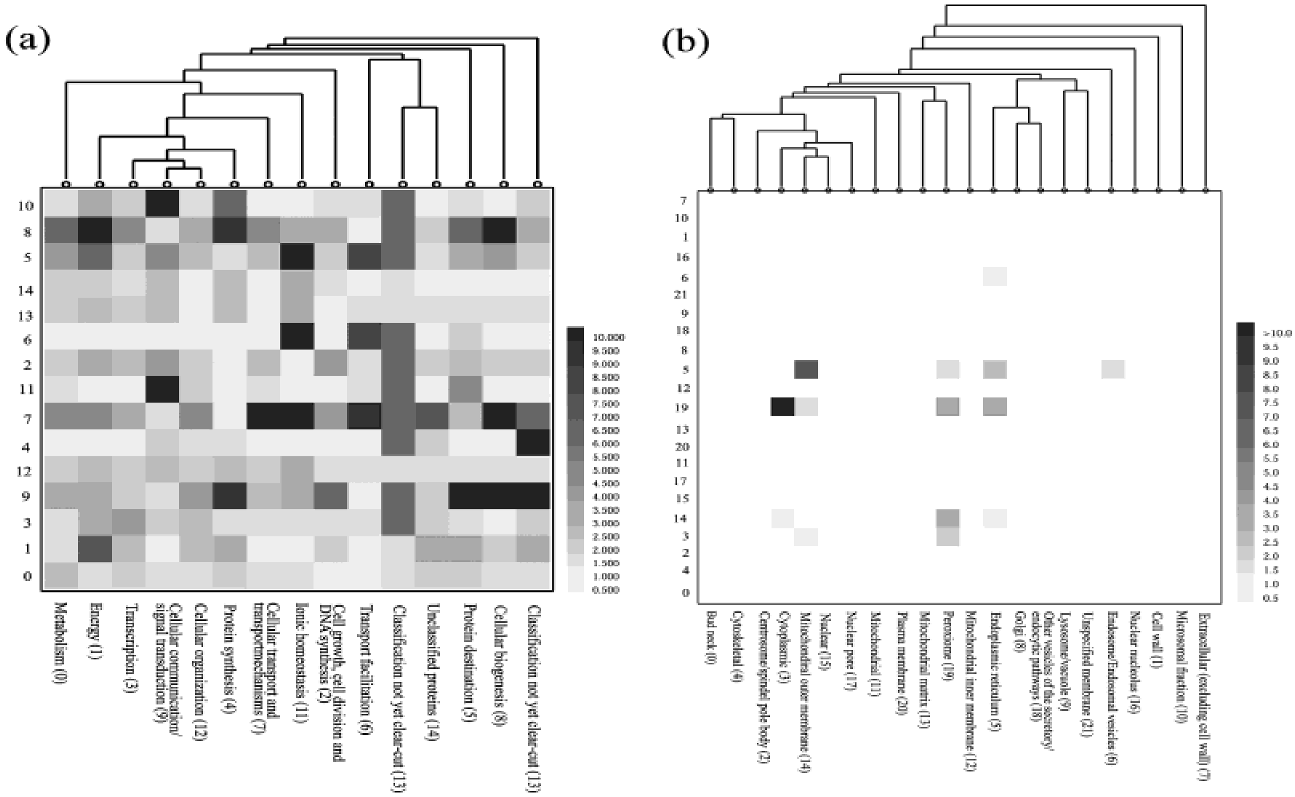


Figure 2. Hierarchical tree and the segregation matrices of the Ito core data based on (a) functional call and (b) subcellular localization. The details of the distance metric are provided in the manuscript.