

# Physical and functional modularity of the protein network in yeast

Thomas Wilhelm<sup>§\*</sup>, Heinz-Peter Nasheuer<sup>†</sup> & Sui Huang<sup>‡</sup>

<sup>§</sup> Institute of Molecular Biotechnology,

Beutenbergstr. 11, D-07745 Jena, Germany

<sup>†</sup> Department of Biochemistry, National University of Ireland,

University Road, Galway, Ireland

<sup>‡</sup> Department of Surgery, Children's Hospital,

Harvard Medical School, Boston, MA 02115, USA

---

\*corresponding author; e-mail: wilhelm@imb-jena.de, phone: +49 3641 656208, fax:  
+49 3641 656191

**Running title:** Modularity of the protein network in yeast

**Abbreviations:**

TAP - Tandem affinity purification

HMS-PCI - High-throughput mass-spectrometric protein complex identification

PN - Protein-protein interaction network

CN - Protein complex interaction network

## Abstract

While protein-protein interactions have been studied largely as a network graph without physicality, here we analyze two protein complex data sets of *Saccharomyces cerevisiae* to relate physical and functional modularity to the network topology. We study for the first time the number of different protein complexes as a function of the protein complex size and find that it follows an exponential decay with a characteristic number of about 7. This reflects the dynamics of complex formation and dissociation in the cell. The analysis of the protein usage by complexes shows an extensive sharing of subunits that is due to the particular organization of the proteome into physical complexes and functional modules. This promiscuity accounts for the high clustering in the protein network graph. Our results underscore the need to include the information contained in observed protein complexes into protein network analyses.

Metabolic and signaling functions as well as global cell behavior arise from the collective action of proteins which engage in physical interactions. Thus, a first step in the functional characterization of the proteome is the identification of protein-protein interactions. This has most exhaustively been achieved for the budding yeast (*S. cerevisiae*) proteome, resulting in large lists of interaction pairs (1,2). This information has allowed the reconstruction of a crude map of the protein interaction network (Fig.1A). Although such network maps are still devoid of any information on dynamics that would allow the simulation of cell behavior (3,4), they have paved the way to the study of global topological properties of molecular networks which have shed light on basic evolutionary and organizational principles (5-7). For instance,

using yeast two-hybrid data it has been suggested that the connectivity distribution  $P(k)$ , i.e. the probability that a protein interacts with  $k$  other proteins, follows a power law and therefore belong to the topology class of scale-free networks (5).

These topology studies relied on the abstract network graphs that were constructed with individual pairs of interactions identified separately, and as such do not represent a physical entity. They do not consider the fact that many protein-protein interactions in the cell take place in dynamic, multi-protein complexes (Fig.1B). Thus, when placing the topology of the protein interaction graph into a physical context, questions automatically arise, as to whether a highly connected protein (a hub in the scale-free network model), would simultaneously interact with all of its partners as denoted in the network graph and, in doing so, form one stable, observable protein complex. Further, one would also like to know how the protein complexes (physical modules) relate to the clusters of highly connected nodes in the network graphs (topological modules). The recent systematic survey of stable protein complexes using high-throughput mass spectrometry of purified tagged yeast proteins now allows us to examine generic aspects of the large-scale properties of complex mediated networks and to address this type of questions (8,9). The data sets are denoted according to the authors as (i) the TAP (tandem affinity purification) (8), and (ii) the HMS-PCI (high-throughput mass-spectrometric protein complex identification) (9) data sets. Their accuracy has been compared with that of other methods of protein-protein interaction detection (1). They have also been used to validate and complement existing yeast interaction datasets and to infer the function of individual proteins

(10). Thus, while the protein complex data has been used to improve functional annotation of individual proteins, the additional, generic information in these datasets, notably the complex size distribution and the pattern of usage of subunits in various complexes has not been studied explicitly. Here we investigate these two generic aspects of the population of yeast protein complexes, identify some characteristic features and propose models to explain them.

## EXPERIMENTAL PROCEDURES

### Stability and concentration of protein complexes

The number  $N_s$  of possible dissociations for a complex consisting of  $s$  proteins is  $N_s = \sum_{i=1}^{s/2} \binom{s}{i} - \binom{s}{s/2}/2$  if  $s$  is even, and  $N_s = \sum_{i=1}^{(s-1)/2} \binom{s}{i}$  if  $s$  is odd. It follows the simple exact result  $N_s = 2N_{s-1} + 1 = 2^{s-1} - 1$ . If all possible dissociations of a given complex occur on average with equal probability it follows the exponential decay of the average lifetime  $\langle \tau_s \rangle$  of a complex with  $s$  proteins:  $\langle \tau_s \rangle \propto N_s^{-1}$ . If the number  $S_s$  of all complexes of size  $s$ ,  $S_s = \sum_i^{n_s} x_{s,i}$  ( $n_s = n_0 \exp(-as)$  is the number of different complexes of size  $s$  (Fig.2), and  $x_{s,i}$  is the number of complexes of species  $i$  (consists of exactly the same type of  $s$  proteins)), is proportional to  $\tau_s$  it follows by using the above equations for  $N_s$  and  $\langle \tau_s \rangle$ :  $S_s/S_{s-1} \cong 0.5$ . If for each complex size  $x_{s,i}$  is normally distributed around  $\langle x_{s,i} \rangle$  it follows by using the equations for  $S_s$  and  $n_s$ :  $\langle x_{s,i} \rangle \exp(-as) / (\langle x_{s-1,i} \rangle \exp(-a(s-1))) \cong 0.5$  and therefore with the experimentally determined characteristic number of different protein complexes  $a = a_{compl.size}$  (Fig.2):  $\langle x_{s,i} \rangle \approx 0.6 \langle x_{s-1,i} \rangle$ . This finding suggests that the mean number of one type of protein complexes of a given size decreases by 40% if the

size is increased by one.

### Construction of protein-protein interaction networks (PN):

#### Three definitions of protein interactions (Fig.1)

A “small *PN*” variant counts only the interactions A-B where protein A as tagged bait catches protein B and protein B as bait catches protein A in the mass spectroscopy protein complex analysis. In doing so one obtains *PNs* with only 193 proteins and 191 interactions for TAP, and 99 proteins and 67 interactions for HMS-PCI. A less stringent “medium *PN*” variant counts all interactions between the tagged bait protein and all of the members of the complex it pulls down as interaction partners. This leads to 1,365 proteins and 3,230 interactions for TAP, and 1,544 proteins and 3,481 interactions for HMS-PCI. The “large *PN*” definition assumes that all proteins participating in a complex interact with each other (directly or indirectly). The intersection between TAP and HMS-PCI data sets yields 0 interactions for the “small” definition, 217 proteins and 191 interactions for “medium”, and 452 proteins and 1,773 interactions for “large” definition. The set union yields 279 proteins, 258 interactions for the “small”, 2,280 proteins, 6,520 interactions for the “medium”, and 2280 proteins, 52,334 interactions for the “large” definition.

#### **Fig.1**

#### Construction of protein complex interaction networks (CN) (Fig.1b)

In this network graph, one node represents a whole protein complex, and two complexes are connected if both contain one (or more) same protein, as proposed (8).

## Measures to characterize protein networks

(i) connectance  $C = 2(\text{number of actual links in the network})/(n(n-1))$ ,  $n$  is the number of nodes (in the  $PN$ , for instance,  $n$  is the protein number); (ii) diameter  $D$  of the largest cluster:  $D$  is the number of links in the shortest path between two nodes, averaged over all pairs of nodes; (iii) clustering index  $cc = \sum_i c_i/n$  with  $c_i = 2k_i/(n_i(n_i-1))$  ( $k_i$  is the number of connections between the  $n_i$  neighbours of node  $i$ ) (11).

## $PN$ and $CN$ are small-world networks

Small-world networks are highly clustered (like regular networks), but have nevertheless a small network diameter (like random networks) (11). Both requirements are fulfilled by the  $PN$  and  $CN$  as can be seen in Tables 1 and 2.

The diameters  $D$  of the  $PN$  and  $CN$  should be compared with that of the corresponding random and regular matrices. A corresponding random matrix is defined as a constructed network, such that it has the same connectance  $C$  as the experimental network, but all links between any two nodes are set by chance: each individual link has a probability of  $1/C$ . Similarly, a corresponding regular matrix is a constructed network that also has the same connectance as the experimental network, but all  $Cn$  links of each node are drawn just to its nearest neighbours, such that a ring-like network structure with a high clustering index appears, cf.(11). Then, for the experimental  $PN$ s (according to the “large” definition) the corresponding random matrix has the following values for the diameter  $D$  of the largest cluster: 2.50 (TAP) and 2.24 (HMS-PCI). The diameter for the corresponding regular matrix of  $PN$  has the values 24.3 (TAP) and 18.4 (HMS-PCI). Accordingly, for the “medium” definition of  $PN$ s one obtains the values 4.40 (TAP), 4.82 (HMS-PCI) for the random matrices, and 125.1 (TAP), 164.5

(HMS-PCI) for the regular matrices. Note that for our “small” definition of  $PN$ s no large connected clusters appear (For the very small connected subgraphs we cannot reasonably attribute statistical features such as “small world”). For the  $CN$ , which do not rely on any definition of protein interactions, one obtains the diameter values 2.26 (TAP) and 2.04 (HMS-PCI) for the corresponding random matrix, and 10.4 (TAP) and 8.0 (HMS-PCI) for the corresponding regular matrix. All values are averages over 20 runs of simulations.

The clustering index of random matrices equals their connectance  $C$ , whereas the clustering index of regular matrices is somewhat below one.

## RESULTS

### The Number of Different Protein Complexes of a Given Size

The analysis of the TAP and HMS-PCI data shows that the number of different protein complexes decays exponentially with the size  $s$  of the protein complex  $f(s) \propto \exp(-as)$  (Fig.2), for the TAP data astonishingly exactly, whereas the HMS-PCI complexes follow this function up to  $s = 15$  but have some more different large complexes ( $a$  is a constant). For TAP and HMS-PCI we find a characteristic number  $s^* = 1/a = 7.3$  and  $6.4$ , respectively. Despite similar average number the HMS-PCI data exhibit more different larger complexes (step in the cumulative graph in Fig.2, around  $s = 15$ ). The exponential decay of the number of different protein complexes with size  $s$  may have implications on the underlying dynamics of complex formation and dissociation.

We propose here a simple model that considers the observed “destabilizing effect” when a given complex grows by one subunit. With increasing size  $s$  of a complex (i.e., containing  $s$  proteins), it has  $N_s$  possible ways of dissociation, where  $N_s = 2^{s-1} - 1$  (for details, see EXPERIMENTAL PROCEDURES). Assuming that  $S_s$ , the number of all complexes (of all species) of size  $s$  is proportional to the average life time, which in turn is inversely proportional to  $N_s$ , it follows that  $S_s/S_{s-1} = 0.5$ . Since  $S_s$  is related to the observed exponential decay  $f(s)$  with the characteristic number  $s^*$ , we can estimate that the average number of complexes with a given composition decreases by 40 % when the complex size increases by one additional subunit (EXPERIMENTAL PROCEDURES). It should now be possible to test this quantitative prediction experimentally in order to validate the physical interpretation of the complex size distribution.

**Fig.2**

The Protein-Protein Interaction Network

To map the physical protein complexes onto a protein-protein interaction graph that represents the topology of the protein network ( $PN$ ), we have to extract the interaction information from the complex data. In contrast to the yeast-two hybrid data, where the elementary experimental finding is a pair that directly translates into a link in the interaction graph, the protein complex data allow various definitions of interaction to build a  $PN$  (see EXPERIMENTAL PROCEDURES). However, the context of a given complex might enable inherently weak, direct physical interactions to take place which would not be found in isolation or in other complexes, e.g. due to the presence of a scaffolding protein in that complex. For instance, while

bait A might not be able to pull out protein B, bait C might pull out a complex that includes A and B which may or may not have direct physical contact. To embrace these scenarios of indirect and scaffold-protein mediated interactions, we use here a “large  $PN$ ” definition that counts “functional interactions” between all proteins participating in a complex (Fig.1C), as it also was suggested (1). Our “medium  $PN$ ” corresponds to the “spoke” model, while the “large  $PN$ ” corresponds to the “matrix” model in a previous study (10). We used the latter, most encompassing  $PN$  definition for further analysis, since we are interested in the observed complexes as entities rather than in the interactions (Table 1, Figs. 2,3). However, similar results with respect to the major network topology characteristics were obtained with the “small” and “medium  $PN$ ” definition (Table 2).

The number of protein interactions  $I$  is an order of magnitude higher than that of networks determined by the combined two-hybrid experiments (5) ( $I = 19,995$  and  $34,112$  vs.  $2,240$  interactions), although the number of proteins involved is smaller ( $1,365$  and  $1,544$  vs.  $1,870$  proteins) (cf. Tables 1,2). Interestingly, we find that the distribution of connectivity  $k$  (number of interaction partners per protein) in the  $PN$  follows an exponential decay, i.e. the  $PN$  is not scale-free as reported for pairwise interaction data (Fig.2). We obtain as the characteristic numbers of interactions per protein,  $k^* = 30$  for TAP, and  $k^* = 47$  for HMS-PCI, thus the average characteristic connectivity of  $PN$  is 38. The larger number of interaction partners for the HMS-PCI data set is consistent with the finding that the HMS-PCI data set contains larger complexes as discussed above. Thus, the number of simultaneous (direct or indirect) physical interaction partners as defined by the coexistence in a pro-

tein complex (under a given culture condition) behaves differently from the number of interaction partners defined by isolated, pair-wise characterization which appears to exhibit a power-law distribution (5).

To measure the extent of modular organization in the large  $PN$  graph we calculated the clustering coefficient  $cc$  which quantifies for a given network the extent of formation of subnetworks (clusters) that are highly interconnected *inter se*. Column 2 of Table 1 shows that  $cc$  of  $PN$  is much higher than the clustering coefficient in corresponding random networks with the same connectance  $C = 2I/(n(n-1))$ , since  $cc_{random} \simeq C_{random} = C_{PN}$ . This quantifies the high modular organization of the cellular protein interaction network.

### The “Null Model”

Since in the large  $PN$  definition all proteins in a complex are considered to interact with each other, the complex as a physical module will necessarily give rise to a maximally connected cluster (=clique) in the network graph. We thus asked whether the partitioning of the proteome into complexes of the observed size distribution alone explains the high clustering. To answer this question we simulated the simplest model, called “null model”. We generated 455 and 487 complexes with the exponential size distribution corresponding to the observed complex size distribution in the TAP and HMS-PCI data, respectively. “Proteins” were randomly drawn (without removing them) from a pool with  $ng$  proteins ( $ng$  is chosen to obtain the same number of “proteins” as in the experimental  $PN$ :  $ng_{TAP} = 1450$ ,  $ng_{HMS-PCI} = 1700$ ). A given

protein can be assigned to more than one complex, but no two same proteins can occur in one complex.). Then an interaction graph is extracted as defined above according to the large  $PN$  scheme and the topology is analyzed. For TAP, this null model yields an exponential distribution for connectivity  $k$  that is similar to the observed one (Fig.2), although the total number of interactions  $I$  in the simulation is higher than in the TAP data (Table 1). In contrast, for HMS-PCI the null model yields a distribution of  $k$  that is clearly steeper than in the observed data, and consistently, the number of interactions  $I$  in the data are higher than predicted by the model (Table 1). This is in line with the notion that the HMS-PCI data contain more larger complexes than a pure exponential size distribution (as assumed for the “null” models), as e.g. shown by the TAP data, would allow.

Interestingly, in both cases the simulated  $cc$  was significantly smaller than in the corresponding experimental  $PN$  ( $cc_{null} = 0.49$  vs.  $cc_{exp} = 0.73$  for TAP; and  $cc_{null} = 0.54$  vs.  $cc_{exp} = 0.70$  for HMS-PCI). Thus, the  $PN$  are strongly clustered, to an extent that cannot be accounted for by the physical arrangement of proteins into complexes that represent cliques in the interaction graph. In other words, the high  $cc$  value of the  $PN$  must be due to higher-level interactions between the physical complexes.

## The Protein Complex Interaction Network

Since the complexes detected by mass spectroscopy are by definition independent entities, such an apparent link between complexes in the network graphs must correspond to the sharing of the same protein by different com-

plexes (Fig.1B). We thus analyzed the topology of the complex-complex interaction network (CN). Fig.3 shows that the connectivity distribution of the CN again follows an exponential decay. A comparison of columns 6 and 8 in Table 1 shows that the simulation gave rise to analogous results as for the PN: the null model yielded more interactions than observed for TAP, and fewer for HMS-PCI. The clustering coefficients for both,  $cc_{exp}$  and  $cc_{null}$  are much higher than the  $cc$  values of the corresponding random networks (Table 1). Again, as for the PN, in the CN the clustering coefficients of the experimentally determined networks,  $cc_{exp}$  (0.52 and 0.54 for TAP and HMS-PCI, respectively) are still considerably higher than the simulated one  $cc_{null}$  (0.30 and 0.28, respectively). The difference  $cc_{exp} - cc_{null}$  is nearly the same for the two network types, PN and CN.

### Fig.3

The higher clustering in both, the experimental PN and CN in comparison to the null model indicate that the latter does not fully determine the PN and CN topology. The finding that even at the higher-level of the CN the experimental cluster coefficient,  $cc_{exp}$  is considerably higher than the simulated one,  $cc_{null}$ , appears to point to a kind of “super-clustering”. In fact, protein complexes are not random aggregates of subunits but represent functional entities that perform specific cellular functions. Moreover, as recently suggested, complexes that perform similar cellular roles and belong to the same functional group (such as cell cycle, mRNA metabolism,transcription, etc.) extensively share proteins (8,9,12,13). Gavin *et al.* (8) proposed 9, and Mewes *et al.* (12) 11 of such functional groups.

## The “Stage One Model”

To account for the bias introduced by the sharing of proteins between functionally related complexes, we extended the null model to a stage one model. Herein, the pool of  $ng$  proteins ( $ng_{1TAP} = 1650$ ,  $ng_{1HMS-PCI} = 1800$ ) from which the complex subunits are drawn is now divided into  $gr$  functional groups of equal size. With a high probability  $pr$  we took the “proteins” for a given complex from the same group to capture the finding that a complex with a certain cellular function contains mostly proteins that have been assigned to the same functional category. The connectivity distribution of the corresponding  $PN$  extracted from the “stage one model” with  $gr = 10$ ,  $pr = 0.9$  is shown in Fig.2 (crosses). In the case of the  $PN$ , when compared to the null model the new model only slightly changes the distribution of  $k$  by shifting the weight to the tail. In the case of the  $CN$ , the stage one model increases the decay of the exponential distribution of connectivity in  $CN$  as compared to the null model (Fig.3) and strongly decreases  $I$  (Table 1). This finding suggests that the increased promiscuity of complexes is not associated with an increase of new links between previously unconnected complexes, but instead, results from the increase of number of links between already connected complexes, reflecting the sharing of multiple proteins.

With  $gr = 10$ ,  $pr = 0.9$  the cluster coefficient for the stage one model,  $cc_{one}$ , is only slightly (but significantly) higher than  $cc_{null}$  but still fails to produce the observed high value of  $cc_{exp}$  of  $PN$  and  $CN$ . However, with higher values of  $pr$  and  $gr$  the clustering coefficient increases; and for  $pr = 0.99$  and  $gr = 100$  the clustering coefficient reaches the experimental values,

$cc_{one} \simeq cc_{exp}$ . Taking into account the combination of functional and spatial cellular compartmentalization  $gr = 100$  may not be an overestimation, since Ho *et al.* discriminated 34 functional and 15 spatial groups (9), and Schwikowski *et al.* discussed 42 functional and 9 spatial groups (13).

**Table 1**

### Comparison with other Protein Interaction Networks

As mentioned above, in contrast to the  $CN$ , the  $PN$  depend on the assumed definition for protein interactions (see EXPERIMENTAL PROCEDURES). Table 2 shows that the  $PN$ s corresponding to our "small" and "medium" definition do also belong to the class of small-world networks. The clustering coefficients  $cc$  of these  $PN$ s are more than one magnitude higher than that of the corresponding random networks  $cc_{crn}$ . Note that the clustering coefficient of random networks equals the connectance of these networks:  $cc_{crn} = C$ . However, it is remarkable that the TAP networks have a higher  $cc$  than the HMS-PCI networks.

For the sake of comparison we also add the analysis of three other protein interaction data sets: (i) Y2H data for yeast protein interactions, as analysed in (5) (data on the website [www.nd.edu/~networks/cell](http://www.nd.edu/~networks/cell)), (ii) the carefully curated yeast protein interaction data set discussed in (2) (data on the DIP website [dip.doe-mbi.ucla.edu](http://dip.doe-mbi.ucla.edu)), and (iii) protein interactions of the human signal transduction network of the TRANSPATH data base ([www.transpath.de](http://www.transpath.de)). Again, these  $PN$ s have much higher clustering coefficients than their corre-

sponding random matrices (Table 2).

It has been claimed that protein networks are of the scale-free type (5), i.e. the distribution of the number of connections per protein should follow a power-law. In contrast, we have shown that both, the distributions for the *PN* ("large" definition, Fig.2), and the *CN* (Fig.3) clearly follow an exponential law  $p(k) \propto \exp(-ak)$ . For all the networks analysed in Table 2, the corresponding distributions are between a power-law and a pure exponential law: all these connectivity distributions follow a stretched exponential distribution:  $p(k) \propto \exp(-ak^b)$ , with  $b < 1$ .

## Table 2

## DISCUSSION

Using the protein complex data in yeast obtained with the TAP and HMS-PCI techniques (8,9), we derived different protein-protein interaction networks. Our interaction definitions yielding the "medium" and "large" protein networks correspond to the recently published "spoke" and "matrix" model, respectively (10). We favour the "matrix" model, because each protein in a given complex interacts physically and/or functionally with each other protein in this complex. Accordingly we studied more in detail our "large" protein interaction network. In agreement with other protein network studies (14,15) we also find the small-world property. As we have shown, this result does not depend on the assumed kind of definition for protein interactions.

In contrast to others (5,14,15), we cannot affirm that protein networks are of the scale-free type. We find that the distributions of the number of connections per protein clearly follow an exponential law, or a stretched exponential law. This may have implications for the evolution of protein networks, because scale-free networks need some preferential attachment to arise, without preferential attachment exponential networks emerge (16). The protein complex networks which do not rely on special definitions, also show the exponential connectivity distribution. Our null model reveals that this is mainly due to the exponential distribution of the number of different protein complexes of a given size. However, in order to explain the high clustering, both in *PNs* and *CNs*, we had to expand the null model. Although the *pr* and *gr* values are somewhat arbitrary, as is the ontological classification of proteins into functional groups, the stage one model reveals an interesting property of protein complexes: The ingredient to be added to the minimal null model to reproduce the high clustering coefficients observed in the *PN* and the *CN*, is the massive overlap of protein subunit usage by the complexes, caused by the use of highly similar combinations of proteins in complexes with similar cellular roles. This extensive promiscuity between complexes is what gives rise to high clustering coefficients in the network topology, and thus to the impression of modularity.

We have shown that the statistical properties of the TAP and HMS-PCI data slightly differ in some details, especially the HMS-PCI data contain more large complexes. This may be due to the fact that the HMS-PCI data were obtained by overexpressing the tagged protein which could have resulted in increased chance of pulling out weakly interacting proteins. Furthermore, in

the tandem affinity purification the complexes are purified in a two-step procedure, in contrast to the one step procedure used by HMS-PCI, which could also contribute to a finding of weaker interactions by HMS-PCI. Interestingly, the perhaps weaker interactions mainly occur in larger protein complexes. Knowledge-based analysis of the two data sets, TAP and HMS-PCI, showed that in HMS-PCI the bait often co-purified complexes of independent origin resulting in larger complexes. In contrast, TAP yielded single complexes in most cases. These complexes often consist of only core complexes with some already biochemically and immunologically characterized subunits or auxiliary proteins missing (data not shown). Consider, for example, the biochemically and immunologically purified complex replication factor C (RFC) which has 5 subunits called RFC1 - RFC5. In TAP, the bait protein RFC2 pulled down RFC3 and RFC4, and one additional protein, EFD1 (8, supplementary data). The same bait protein, RFC2, also co-purified RFC3 and RFC4 in HMS-PCI, but pulled down 13 additional proteins as well (9, supplementary data). These questions require to be examined more carefully in future studies. However, with our simple null and stage one model we can reproduce the main features of the underlying protein networks. Further refinements of these models can be done for more consistent future data.

Our results show that the graph-theoretical analysis of clustering and modularity in the topology of protein interaction networks needs to take into account the observed physical modules (complexes) and their particular organization into functional modules and higher-order (complex-complex) networks by the shared usage of proteins. These aspects are lost in the usual graph representation of protein networks. The importance of the analysis of

physical protein complexes is also underscored by the demonstration of the exponential distribution of the number of different complexes of a given size that reflects the physicochemical dynamics of complex formation and dissociation. This has been shown already for the exponential distribution of the number of domains in proteins (17). Higher quality, exhaustive protein complex data in the near future will allow one to translate the topological maps of biochemical networks that contain potential interactions into complexes defined by actual physical interactions.

We thank A. Beyer, F. Grosse and J. Sühnel for critical reading of the manuscript and an anonymous referee for valuable comments. This work was supported by BMBF grants.

## REFERENCES

1. von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.
2. Deane, C. M., Salwiński, L., Xenarios, I., and Eisenberg, D. (2002) Protein interactions. Two methods for assessment of the reliability of high throughput observations. *Mol.Cell.Prot.* **1**, 349-356.
3. Huang, S. (2001) Genomics, complexity and drug discovery: insights from Boolean network models of cellular regulation. *Pharmacogenomics* **2**, 203-222.
4. Smith, A. E., Slepchenko, B. M., Schaff, J. C., Loew, L. M., and

- Macara, I. G. (2002) Systems analysis of ran transport. *Science* **295**, 488-491.
5. Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001) Lethality and centrality in protein networks. *Nature* **411**, 41-42.
  6. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555.
  7. Wagner, A., and Fell, D. (2001) The small world inside large metabolic networks. *Proc.Roy.Soc.London, B* **268**, 1803-1810.
  8. Gavin, A.-C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147.
  9. Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-183.
  10. Bader, G.D., and Hogue, W.V. (2002) Analyzing yeast protein - protein interaction data obtained from different sources. *Nature Biotechnology* **20**, 991-997.
  11. Watts, D. J., and Strogatz, S. H. (1998) Collective dynamics of small-world networks. *Nature* **393**, 440-442.
  12. Mewes, H. W. *et al.* (1997) Overview of the yeast genome. *Nature* **387** (Suppl.) 9.

13. Schwikowski, B., Uetz, P., and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nature Biotechnology* **18**, 1257-1261.
14. Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283-1292.
15. Snel, B., Bork, P., and Huynen, M. A. (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci.* **99**, 5890-5895.
16. Barabási, A.-L., and Albert, R. (1999) Emergence of scaling in random networks. *Science* **286**, 509-512.
17. Koonin, E. V., Wolf, Y.I., and Karev, G.P. (2002) The structure of the protein universe and genome evolution. *Nature* **420**, 218-223.

## Figure Legends

**Fig. 1.** Protein networks. (A) Binary protein interactions. (B) Protein complexes elucidated with TAP and HMS-PCI (highlighted by a circle). In the corresponding protein complex network (CN) two protein complexes are connected if they share one or more proteins. Dashed lines indicate shared proteins, solid lines indicate the resulting interactions between protein complexes. (C) Protein interaction network (PN) derived from complex data in B according to the “large” definition, where all possible interactions within the protein complexes are considered.

**Fig. 2.** Distribution of the number of different protein complexes of a given size  $s$  (open circles) and of the number of  $k$  connections per protein (solid circles). Triangles and crosses represent the connectivity distributions of the “null” and “stage one model”, respectively. Lines are fitting results with the exponential function  $f(x) \propto \exp(-ax)$  ( $x$  is  $s$  or  $k$ ): (A) TAP ( $a_{compl.size} = .137$ ,  $a_{PN} = .033$ ,  $a_{null} = .022$ ,  $a_{one} = .026$ ), (B) HMS-PCI ( $a_{compl.size} = .156$ ,  $a_{PN} = .021$ ,  $a_{null} = .030$ ,  $a_{one} = .033$ ).

**Fig. 3.** Distribution of the number of  $k$  connections per protein complex in the complex-complex interaction network (solid circles) and the corresponding “null” (triangles) and “stage one model” (crosses). Lines are fitting results with the exponential function  $f(k) \propto \exp(-ak)$ : (A) TAP ( $a_{CN} = .049$ ,  $a_{null} = .048$ ,  $a_{one} = .070$ ), (B) HMS-PCI ( $a_{CN} = .032$ ,  $a_{null} = .056$ ,  $a_{one} = .083$ ).

**Table 1**

Statistics of experimental and simulated networks

measure	protein network				complex network			
	PN	PN(LCL)	“null”	“one”	CN	CN(LCL)	“null”	“one”
<b>TAP</b>								
nodes $n$	1,365	1,250	1,360	1,360	455	412	455	455
interactions $I$	19,995	19,815	34,000	30,000	4,312	4,306	5,200	3,700
connectance $C$	0.02	0.03	0.04	0.03	0.04	0.05	0.05	0.04
clustering $cc$	0.73	0.74	0.49	0.54	0.52	0.57	0.30	0.34
diameter $D$		2.85				2.63		
longest path		7				7		
<b>HMS-PCI</b>								
nodes $n$	1,544	1,501	1,540	1,530	487	469	487	487
interactions $I$	34,112	34,076	29,000	26,000	7,368	7,368	4,800	3,300
connectance $C$	0.03	0.03	0.02	0.02	0.06	0.07	0.04	0.03
clustering $cc$	0.70	0.71	0.54	0.58	0.54	0.56	0.28	0.31
diameter $D$		2.57				2.34		
longest path		6				6		

PN and  $CN$  denote the experimental protein-protein (“large” definition) and complex-complex interaction networks. “null” and “one” denote the simulated networks according to the “null” and “stage one model” (The simulation results are averaged values over 20 runs. The standard deviations are 10, 4000, and 0.01

for  $n$ ,  $I$ , and  $cc$ , respectively).  $LCL$  denotes the largest connected cluster,  $D$  its diameter (i.e. the number of links in the shortest path between two nodes, averaged over all pairs of nodes). Analysis shows that  $PN$  and  $CN$  belong to the class of small-world networks (see EXPERIMENTAL PROCEDURES).

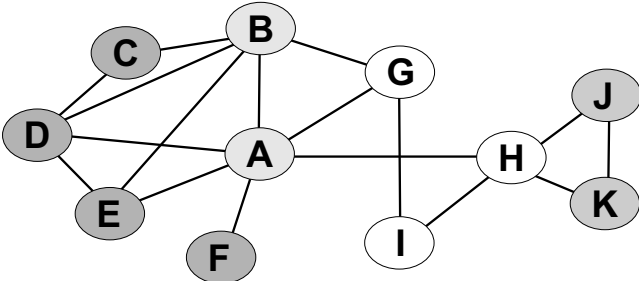
**Table 2**

Statistics of protein interaction networks (PN)

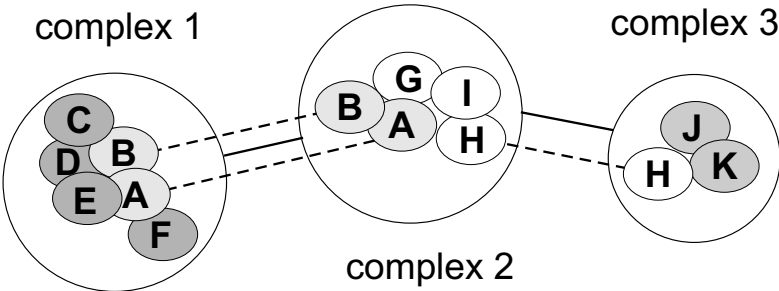
measure	TAP		HMS-PCI		other data sets		
	“small”	“medium”	“small”	“medium”	Y2H	DIP	TP
nodes $n$	193(15)	1,365(1,250)	99(7)	1,544(1,501)	1,870	1,788	434
interactions $I$	191(38)	3,230(3,150)	67(7)	3,481(3,456)	2,240	3,003	868
connectance $C$	0.01(0.36)	0.003(0.004)	0.01(0.33)	0.003(0.003)	0.001	0.002	0.009
clustering $cc$	0.248(0.66)	0.216(0.233)	0.071(0)	0.048(0.049)	0.068	0.188	0.054
diameter $D$	(1.94)	(4.93)	(1.81)	(4.41)			
longest path	(4)	(12)	(3)	(11)			
stretch parameter $b$	0.78	0.48	0.65	0.34	0.34	0.53	0.55

“small” and “medium” denote the corresponding definitions of protein interactions (see EXPERIMENTAL PROCEDURES). In parentheses the corresponding values of the largest connected cluster are given. Y2H denote the combined yeast protein interaction data as analysed in (5). DIP denotes the yeast data set discussed in (2), and TP stands for the human TRANSPATH data base. The stretch parameter  $b$  follows from nonlinear regression of the cumulative connectivity distribution with  $p(k) \propto \exp(-ak^b)$ .

**A Network graph constructed from interaction pairs**



**B Observed complexes and complex network (CN)**



**C Network (PN) derived from complex data**

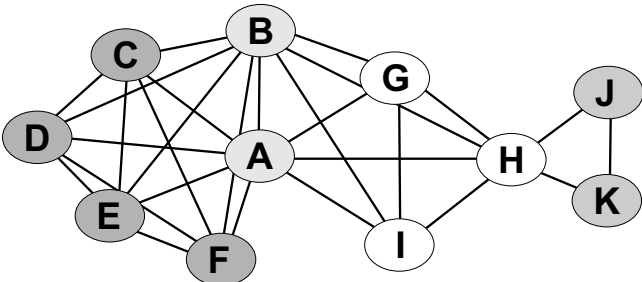


Fig.1

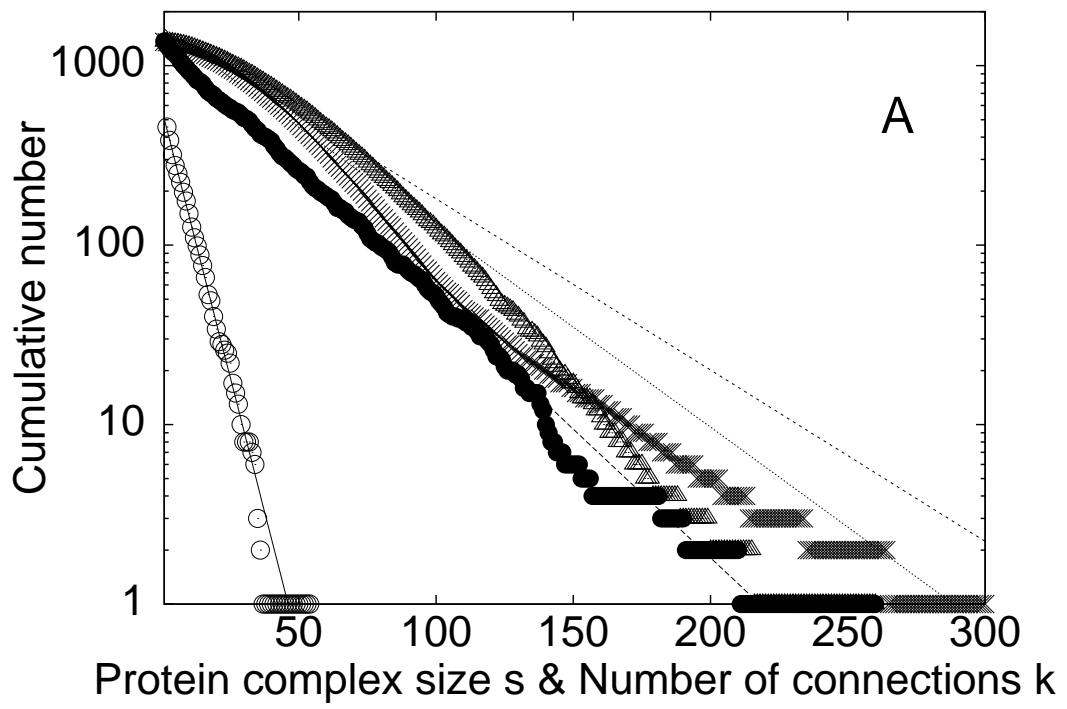


Fig.2a

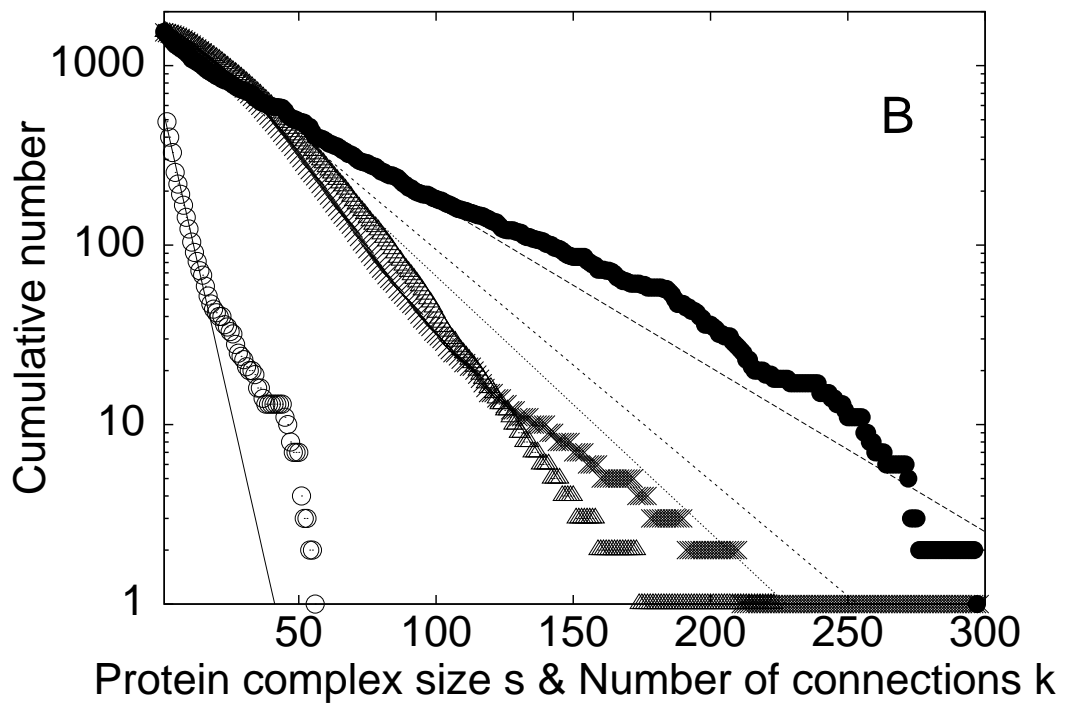


Fig.2b

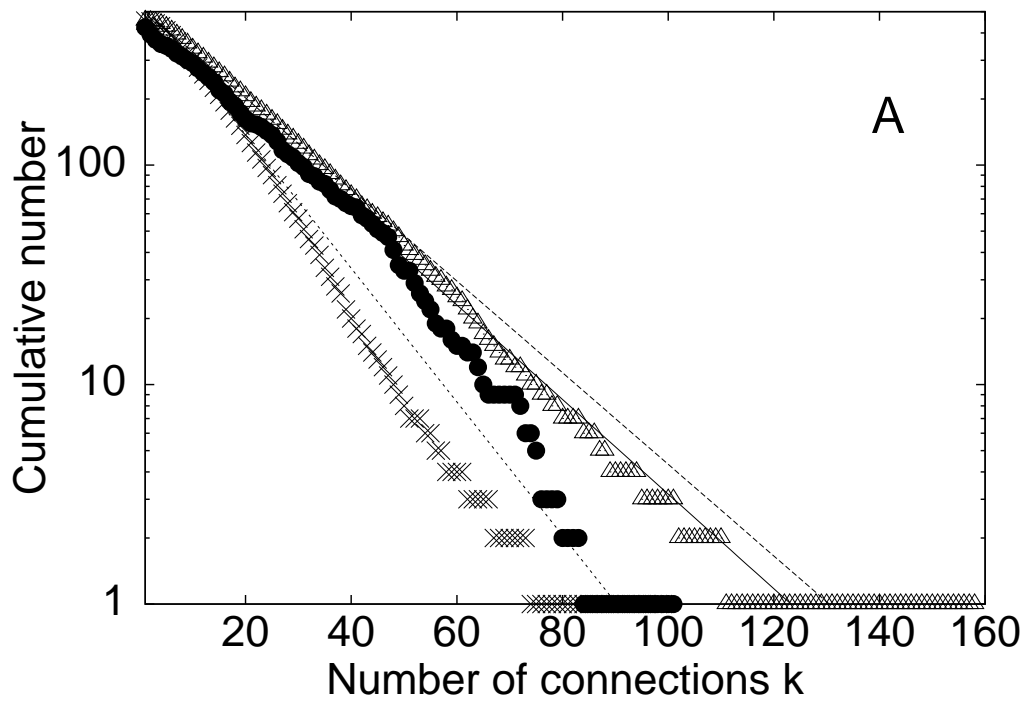


Fig.3a

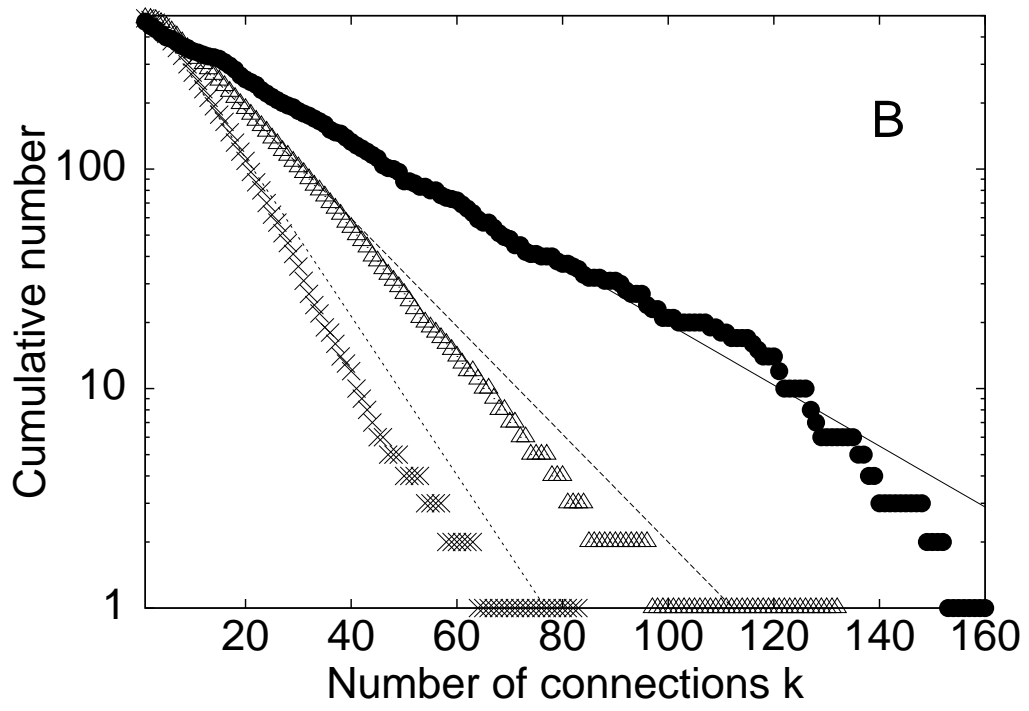


Fig.3b