

REGULAR ARTICLE

Peeling the yeast protein network

Stefan Wuchty and Eivind Almaas

Department of Physics, University of Notre Dame, Notre Dame, IN, USA

A set of highly connected proteins (or hubs) plays an important role for the integrity of the protein interaction network of *Saccharomyces cerevisiae* by connecting the network's intrinsic modules [1, 2]. The importance of the hubs' central placement is further confirmed by their propensity to be lethal. However, although highly emphasized, little is known about the topological coherence among the hubs. Applying a core decomposition method which allows us to identify the inherent layer structure of the protein interaction network, we find that the probability of nodes both being essential and evolutionary conserved successively increases toward the innermost cores. While connectivity alone is often not a sufficient criterion to assess a protein's functional, evolutionary and topological relevance, we classify nodes as globally and locally central depending on their appearance in the inner or outer cores. The observation that globally central proteins participate in a substantial number of protein complexes which display an elevated degree of evolutionary conservation allows us to hypothesize that globally central proteins serve as the evolutionary backbone of the proteome. Even though protein interaction data are extensively flawed, we find that our results are very robust against inaccurately determined protein interactions.

Received: May 4, 2004
Revised: July 12, 2004
Accepted: July 14, 2004

Keywords:

Evolutionary backbone / Local and global centrality / Protein cores

1 Introduction

The available protein-protein interaction network of *Saccharomyces cerevisiae* is a result of large-scale efforts chiefly using high-throughput techniques [1]. The majority of the observed interactions occurs among proteins affiliated within the same functional classes, reflecting a considerable cohesiveness among members of protein complexes. However, a large number of interactions exists between the protein members of functional groups [2, 3]. As a result, this complex network is described as the accumulation of discernible, yet topologically overlapping, functional modules where tight clusters of proteins are connected into larger less cohesive groups [4]. Apparently, networks featuring such functional modules are observed in almost all types of biological systems [5, 6] where a small subset of well-connected nodes, or hubs, plays the important role of linking the network's modules [7, 8]. The importance of the hubs' placement in the network is further confirmed by their

heightened tendency to be lethal [9]. However, little is known about the topological coherence among these highly connected proteins. Here, we employ a method for the decomposition of a network into k -cores [10] (see Fig. 1), allowing us to classify nodes simultaneously by both their connectivity and central placement in the network. These subgraphs emerge through a recursive removal of all nodes with connectivity less than k (see Section 2). Nesting toward the innermost k -core, this procedure allows us to identify and systematically investigate layers of the network. Although the decomposition into k -cores can potentially break the network into disconnected parts, we observe that the yeast protein interaction network keeps its integrity. This method has successfully been applied to protein interaction networks of yeast to determinate and visualize peptide recognition modules and other protein complexes [11, 12]. Yet, a systematic analysis and interpretation of k -core subgraphs is absent.

Although highly emphasized in current literature, we find that connectivity alone is not a sufficient criterion to assess a protein's topological role. In fact, the local neighborhood embedding a particular hub is vital for its potential to be essential and evolutionary conserved. While we observe that the degree of a sparsely interacting protein reflects its topological role reasonably well, we find that the affiliation to

Correspondence: Dr. Eivind Almaas, Department of Physics, 225 Nieuwland Science Hall, University of Notre Dame, Notre Dame, IN 46556, USA

E-mail: almaas.1@nd.edu

Fax: +1-574-631-5259

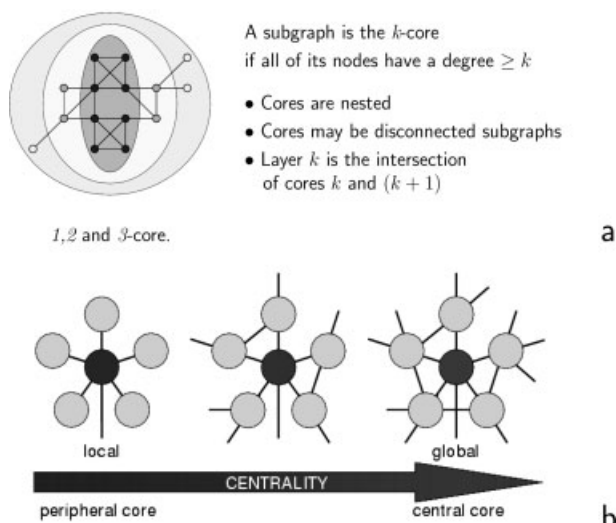


Figure 1. (a) Schematic description of the k -core decomposition of a network. The 1-core consists of all the nodes while the 3-core only contains the nodes on dark gray background. The 2-layer consists of the four nodes on white background. (b) Local vs. global centrality. Interpreted as its importance, the centrality of a node is related to its degree and network neighborhood. A hub that is only a member of the outer k -cores is locally central (top-left), while nodes (not necessarily the biggest hubs) being members of the innermost cores are globally central (top-right).

the innermost k -cores indicates the central placement of a protein significantly better. So, we define highly connected proteins which are merely members of the outer k -cores to be locally central. In turn, proteins in the innermost k -cores, which are not necessarily among the highest connected ones, are defined to be globally central (Fig. 1). These definitions allow us a clear-cut assessment of centrality, which is not offered by the utilization of the node's degree alone. We demonstrate that the group of proteins appearing in the central cores displays a significant excess retention of lethality and evolutionary conservation. Furthermore, such proteins participate in a substantial number of protein complexes. The observation that these complexes largely are fully evolutionary conserved allows us to suggest that the globally central proteins constitute the evolutionary backbone of the yeast proteome. Although the accuracy of yeast protein interactions and orthologs extensively suffers from high rates of false signals we find that our results are very robust in the presence of high levels of noise.

2 Materials and methods

2.1 Protein interactions

The DIP database [13] provides a set of manually curated protein-protein interactions in the organism *S. cerevisiae*. The current version contains 3677 proteins involved in

11 249 interactions derived from combined, nonoverlapping data obtained mostly by systematic two-hybrid analysis.

2.2 Determination of k -cores

The k -core of a graph is defined as the maximum subgraph if every node has at least k links (Fig. 1a). Determining a k -core by iteratively pruning all nodes with a degree lower than k and their incident links, we repeatedly apply the following algorithm: (i) sort nodes according to their present degree and (ii) remove the nodes with degree lower than k . The remaining (sub-)graph is called the k -core of the network [10, 14].

2.3 Assignment of orthologs

The InParanoid database [15] provides orthologous sequence information for *S. cerevisiae* and the complete protein sets of *Homo sapiens*, *Drosophila melanogaster*, *Mus musculus*, *Caenorhabditis elegans*, and *Arabidopsis thaliana*. Utilizing all-versus-all BLASTP searches in protein sets of two species, sequence pairs with mutually best scores are selected as central ortholog pairs. Proteins of both species which show an elevated degree of homology are clustered around these central pairs, a procedure that forms orthologous groups. The quality of the clustering is finally assessed by a standard bootstrap procedure. The central ortholog sequence pair which provides a bootstrap confidence level of 100% is considered as the real orthologous relation while proteins with a lower level of confidence are considered as their paralogs. In our study, we only selected the central ortholog sequence pairs of each group, resulting in 2251 yeast proteins with orthologs in *H. sapiens*, 2149 in *A. thaliana*, 1874 in *C. elegans*, 2046 in *M. musculus* and 2040 in *D. melanogaster*. Additionally, we considered a cross section list of the 704 yeast proteins having an ortholog in each of these higher eukaryotes.

2.4 Assignment of lethality

The BioKnowledge library [16] is created by scans of the experimental literature to provide a comprehensive list of (non-)essential proteins. Of the *S. cerevisiae* proteins appearing in the interaction network, 810 are assigned to be essential while 2704 are considered nonessential.

2.5 Excess retention

The excess retention of proteins with property A (e.g. being (non-)essential or having an ortholog in a higher eukaryote) in the k -core (E_k^A), is the degree to which proteins from this group are over- or underrepresented relative to the original network. First, the fraction of nodes with property A in the full (1-core) network of N nodes is $E^A = N^A/N$. Similarly, for the k -core with a total of N_k nodes, the fraction of proteins

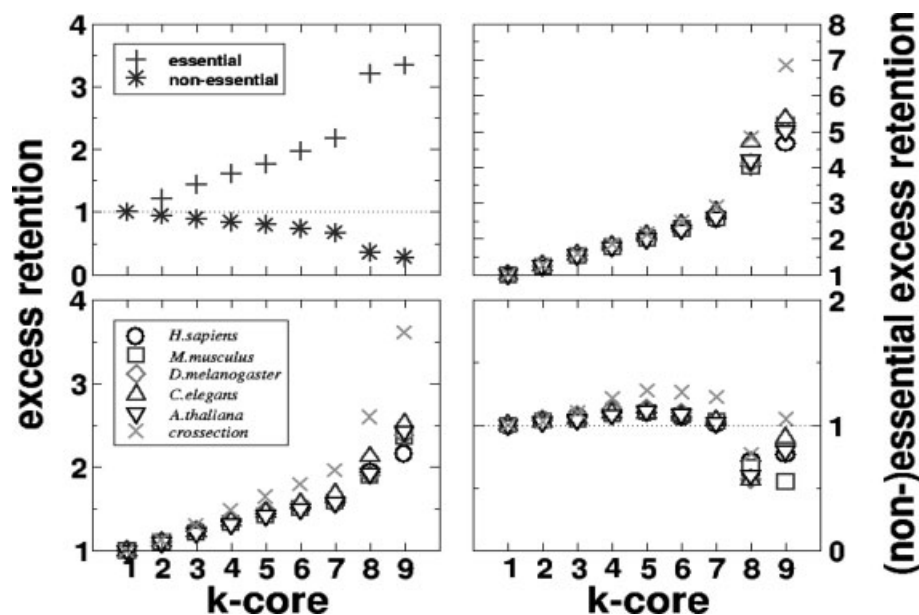


Figure 2. Excess retention of orthologous and (non-)essential proteins in k -cores 1–9 of the yeast protein interaction network. (a) The excess retention of (non-)essential proteins displays a marked increase (decrease) with the k -cores. (b) Analogously, the excess retention of proteins with orthologs in either *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, and *A. thaliana* increases significantly with k . This trend is enhanced in the cross section, which represents the set of proteins with an ortholog in all the five higher eukaryotes. (c), (d) Combining the analysis of (a) and (b), the excess retention of k -core proteins simultaneously having orthologs in a higher eukaryote and being classified as essential (c) (nonessential (d)) shows a clear increase (no trend).

with property A is defined as $e_k^A = n_k^A/N_k$. Hence, the excess retention of nodes with property A in the k -core is given by $ER_k^A = e_k^A/E^A$ [17].

2.6 Protein complexes

In a recent study [18], global mRNA expression patterns [19] were used to analyze a compilation of protein complex information [5, 6, 20]. In this analysis, 428 complex centers were identified in which the subunits are highly co-expressed. Furthermore, the topological proximity of these proteins is further indicated by the same deletion phenotype, identical functional classification and cellular localization. Each complex center is surrounded by a functionally mixed group of proteins which represent spurious and likely short-lived attachments.

3 Results

The protein interaction network of *S. cerevisiae*, consisting of 3677 different proteins participating in 11 249 interactions compiled from the DIP database [13] has 35 disconnected

protein groups and one giant component embracing more than 98% of all the proteins. Applying the k -core decomposition method, we determined the existence of nine consecutive k -cores, which mostly keep the overall statistical properties of the network unchanged (see supplementary material).

The observation that highly interacting proteins have an increased tendency to be lethal [3] should be well reflected in the central k -cores since they only contain prolific proteins. However, the fact that essential proteins on average accumulate only slightly more interactions

than their nonessential counterparts [21] renders the question of k -core lethality nontrivial. Utilizing data from the BioKnowledge library [16], we investigate the possibility that central k -core proteins are preferentially lethal by identifying the fraction of lethal proteins in each k -core. If topology and lethality are uncorrelated, the ratio of the k -core dependent to the expected number of lethal proteins (the k -core excess retention [17]) would be unity (see Section 2). Instead, we find a systematic excess retention of lethal proteins (Fig. 2a) which is threefold for the innermost cores while non-essential proteins appear diluted. To evaluate the influence of the local neighborhoods on this finding, we first ranked the nodes solely according to their degree before we successively split this list of nodes into nine bins (“cores”) with sizes identical to those of the k -cores (see supplementary material). While we find that the resulting curves of the excess retention are largely similar we observe a reasonable difference for globally central proteins.

The accumulation of essential proteins in the innermost cores suggests that they are evolutionary conserved as well, since yeast proteins organized in cohesive interaction patterns are conserved to a significantly high extent [22]. Using the InParanoid database [15] to identify the central ortholog pairs between protein sequences of *S. cerevisiae* and the five higher eukaryotes *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, and *A. thaliana*, we find that the innermost cores display more than a twofold excess retention of orthologs (Fig. 2b). Furthermore, the k -core excess retention of orthologous and (non-)essential proteins is not independent: Figure 2c reveals a fivefold excess retention in the central core of proteins which are simultaneously lethal and orthologous. In contrast, a clear trend for proteins which are both nonessential and evolutionary conserved is absent (Fig. 2d). Focusing on the set of proteins which have orthologs in all higher eukaryotes (the ‘cross section’), we find the initial

Table 1. List of the five most connected proteins in the four innermost core-layers (the intersection of two consecutive k -cores) using their degree in the full network

k -core	Protein	Description	k	l	H	M	F	W	A
9	KC21	Casein kinase subunit	62		*	*	*	*	*
9	YMT9	rRNA processing	50	*	*	*	*	*	*
9	IF6	Translation init. fact. 6	44	*	*	*	*	*	*
9	NOP2	Nucleolar protein	44	*	*	*	*	*	*
9	YK61	Unknown	38	*	*	*	*	*	*
8	IMA1	RNApol I suppressor	170	*	*	*	*	*	*
8	N116	Nuclear pore transporter	121	*	*	*		*	*
8	H4	Histone H4	68		*	*	*	*	*
8	YJZ2	mRNA splicing factor	60		*	*	*	*	*
8	PR06	Pre-mRNA splicing factor	57	*	*	*	*	*	*
7	JSN1	Nucleolar mRNA transporter	225						
7	AT14	ATP synthase subunit	111						
7	TEM1	GTP-binding protein	98	*					*
7	SRB4	RNApol. B suppressor	94	*					
7	TF2B	Transcript. init. factor 2B	89	*	*	*	*	*	*
6	YKA2	Endosomal protein	34		*	*	*	*	*
6	RM09	60S ribosomal protein	29		*	*	*	*	*
6	YNJ2	Unknown	25		*	*	*	*	*
6	RHO1	Rho protein	25	*		*	*	*	*
6	SC17	Vesicular fusion protein	25	*	*	*	*	*	*

Notably, the central k -core is not populated by the largest hubs (column k). The globally central nodes carry evolutionary significance, classified as being both lethal (col. l) and having orthologs in the higher eukaryotes *H. sapiens* (col. H), *M. musculus* (M), *D. melanogaster* (F), *C. elegans* (W) and *A. thaliana* (A).

trends enhanced (Figs. 2b, c). We observe that the area toward the outermost cores largely overlaps with our 'null-hypothesis' of binning proteins according to their ranked degree. However, the excess retention in the two innermost k -cores is clearly enhanced, demonstrating that degree alone is insufficient to appraise the biological importance of a protein (node) in a highly clustered area of the network.

Although our results appear promising, we have to evaluate the effect of incorrectly labeled protein interactions. The significantly inaccurate determination of protein interactions (up to 50–90% false-positives and false-negatives) is a well known and serious problem of the high-throughput Y2H method [4, 23]. To estimate the potential influence of incomplete and noisy protein interaction data on our findings, we added (removed) up to 75% of interactions between randomly selected protein pairs, thereby mimicking false-positives (false-negatives). Similarly, simulating the presence of false-positive (false-negative) ortholog and essential signals we randomly increased (decreased) the original sets of respective proteins by up to 75%. In each case, we generated 1000 different realizations of removal (addition) and repeated our assessment of essential and orthologous k -core excess retention. We find that the basic trends remain qualitatively unaltered, allowing us to conclude that the uncovered correlations are largely unaffected by severe data incompleteness (see supplementary material for details). The observation that our findings are largely unaffected is in notable contrast

to the determination of hubs, which dramatically vary between the different protein interaction data sets [24].

Which proteins are in the central cores? We find that the proteins which populate the innermost cores are considerably different to sets of proteins we obtain by ranking them by their degree and filling them successively into bins of equal size of the corresponding core. The list of the five most connected proteins in the four innermost core-layers (the k -layer is the intersection of cores k and $(k + 1)$) clearly supports our initial assumption that a protein's connectivity in the full network is not tantamount to ending up in the central core (see Table 1). For example, the largest hub of the protein interaction network, the nucleolar mRNA transporter JSN1 with 225 interactions, only makes it to the 7-core layer. In contrast, the rRNA processing YTM9 protein with only 50 interactions, a mere No. 31 on the ranked list of the most connected hubs, ends up in the main core. The stepwise allocation of nodes into k -cores (or layers) enables us to systematically and unambiguously determine a node's centrality. Hence, the connectedness of a node and its neighbors, as measured by the k -core decomposition method, determines its topological importance: Consider a set of nodes with only a single link connected to a central node (Fig. 1b top-left). As the majority of this hub's neighbors has degree one, it will be removed in the early stages of the k -core decomposition process. On the other hand, a well-connected node embedded in a web of other well-connected nodes will

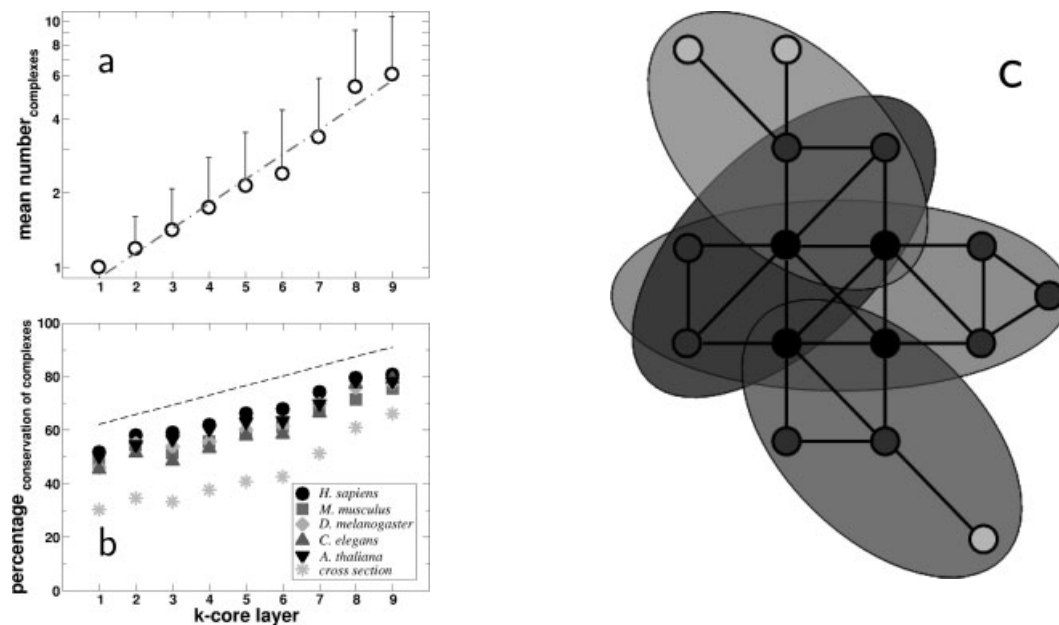


Figure 3. Globally central proteins and protein complexes. (a) For each k -core layer, we calculate the average number of complexes the proteins take part in. (b) Likewise, we determined the mean percentage of orthologous conservation of complexes whose proteins participate in the respective k -core layers. The increasing trends suggest that the globally central proteins are found in the overlapping region of the complexes, as illustrated in (c). The four shaded ellipses indicate protein complexes, the light gray, the medium gray and the black nodes are members of the 1-core, 2-core, and 3-core layers, respectively.

likely become a member of one of the central-most k -cores (Fig. 1b top-right). Henceforth, we define highly connected proteins which are merely members of the outer k -cores to be locally central. In turn, proteins in the innermost k -cores, which not necessarily are among the highest connected ones, are defined to be globally central, allowing a clear-cut assessment of centrality. The biological significance of the globally central proteins is also indicated by Fig. 2 and Table 1, where these proteins are simultaneously lethal and evolutionary conserved in the eukaryotes *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, and *A. thaliana*. This suggests that a combination of evolutionary retention and topological placement (as selected by a chosen core) might be used to evaluate the quality of a protein's interaction data.

Many important cellular functions are maintained by protein complexes, acting as molecular machines with varying size and temporal stability. We have already seen that globally central proteins are simultaneously prolific, essential, and evolutionary conserved. We argue that the centrality of these proteins is caused by their particular role in the protein complexes: In a recent study [18], protein complexes were discovered to feature centers of highly co-expressed proteins which mostly display the same deletion phenotype. Henceforth, we calculate the number of complexes a protein is a member of as function of the k -core layers (Fig. 3a). The globally central proteins participate in the largest number of distinct complex centers, suggesting that their immediate network neighborhood, and consequently their centrality, is controlled by these complexes. Il-

lustrated in Fig. 3c, the globally central proteins (dark nodes) are tightly connected to the four different complexes (colored ellipses) like the spoke of a wheel. The noncentral proteins are members of at most 1 complex.

In the light of Fig. 3, the strong inter-dependence of essentiality and orthology (Fig. 2) suggests that the evolutionary importance of the globally central proteins should be reflected in their complexes: layerwise, we determined the mean percentage of orthologous conservation of complexes whose proteins participate in the respective k -core layers. Utilizing the orthologs in the eukaryotes *H. sapiens*, *M. musculus*, *C. elegans*, *D. melanogaster*, and *A. thaliana*, we find an increasing degree of complex conservation toward the innermost cores (Fig. 3b). Since the orthologous globally central proteins are also members of many different complex centers, this finding suggests that the globally central proteins might constitute a putative evolutionary backbone of the proteome.

4 Discussion

The k -core analysis of a network, as exemplified by the decomposition of the protein interaction network of *S. cerevisiae*, offers new information about a network's large-scale organization. Although highly emphasized in the literature, we have demonstrated that connectivity alone is not a sufficient criterion to assess a protein's functional and topological

relevance. The k -core method poses a systematic way to consider the local vs. global significance of a protein, as evidenced by the elevated probabilities of globally central proteins to be both evolutionary conserved and essential to the survival of the organism. Furthermore, the globally central proteins participate in a substantial number of complexes featuring co-expressed constituents. These results suggest that the globally central proteins serve as topological ‘crystallization nuclei’ for protein complexes. The tendency of globally central proteins to be evolutionary conserved suggests that the k -cores both possibly uncover a proteome backbone. Either these highly conserved proteins are ancient and stopped giving rise to any evolutionary innovations or they still might act as nested kernels for the emergence of most biological functions by participating in more than one complex. The latter observation has interesting implications for our understanding of the topological construction of protein clusters: The appearance of such overlapping clusters renders the application of nonoverlapping cluster algorithms for the detection of modules and putative protein complexes limited. However, we still find that globally central proteins act as linkers to facilitate the integration of separate complexes without actually being part of them. Such tasks may be carried out by proteins which do not have to be simultaneously co-expressed with the constituents of the protein complexes they connect. The advent of time-resolved expression data will help to uncover these roles proteins might potentially play in the topological arrangement and management of protein complexes.

Erroneous protein-protein interaction data of yeast seriously jeopardize the strength of our results. Although recent investigations uncovered startling false-positive and false-negative error rates ranging from 50 to 90%, our assessment uncovers surprisingly robust trends. Previous perturbation analyses reported that real networks are robust against random failure of nodes, while they suffer severe damage upon targeted attack of the network’s hubs [25]. While these observations resulted from node-based perturbations, our error analysis focuses on random events on interaction level. Although we observe quantitative changes, extreme false-positive and false-negative error rates of up to 75% did not force the networks to lose its initial characteristic that globally central nodes are predominantly conserved in evolution. Notably, the topology of the underlying network compensates such severe perturbations. This observation is not only a convincing proof of our concept but also underlines the special design of the web these proteins are embedded in.

5 References

- [1] Von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.*, *Nature* 2003, 417, 399–403.
- [2] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. *et al.*, *Nature* 2002, 415, 180–183.
- [3] Gavin, A., Bösch, M., Krause, R., Grandi, P. *et al.*, *Nature* 2002, 415, 141–147.
- [4] Rives, A., Galitski, T., *Proc. Natl. Acad. Sci. USA* 2003, 100, 1128–1133.
- [5] Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., Barabási, A., *Science* 2002, 297, 1551–1555.
- [6] Holme, P., Huss, M., Jeong, H., *Bioinformatics* 2003, 19, 532–538.
- [7] Newman, M., *Phys. Rev. Lett.* 2002, 89, 208701.
- [8] Newman, M., *Phys. Rev. E* 2003, 67, 026126.
- [9] Jeong, H., Mason, S., Barabási, A.-L., Oltvai, Z., *Nature* 2001, 411, 41–42.
- [10] Seidman, S., *Social Networks* 1983, 5, 269–287.
- [11] Tong, A., Drees, B., Nardelli, G., Bader, G. *et al.*, *Science* 2002, 295, 321–324.
- [12] Bader, G., Hogue, C., *Nature Biotechnol.* 2002, 20, 991–997.
- [13] Xenarios, I., Salwinski, L., Duan, X., Higney, P. *et al.*, *Nucleic Acids Res.* 2002, 30, 303–305.
- [14] Batagelj, V., Zaveršnik, M., *An $o(m)$ Algorithm for Cores Decomposition of Networks*, University of Ljubljana, preprint series Vol. 40, 799, <http://vlado.fmf.uni-lj.si/pub/~preprint/imfm0798.pdf> 2002.
- [15] Remm, M., Storm, C., Sonnhammer, E., *J. Mol. Biol.* 2001, 314, 1041–1052.
- [16] Costanzo, M., Crawford, M., Hirschmann, J., Kranz, J. *et al.*, *Nucleic Acids Res.* 2001, 29, 7579.
- [17] Wuchty, S., *Genome Res.* 2004, 14, 1310–1314.
- [18] Dezsó, Z., Oltvai, Z., Barabási, A.-L., *Genome Res.* 2004, 13, 2450–2454.
- [19] Hughes, T., Marten, M., Jones, A., Roberts, C. *et al.*, *Cell* 2000, 102, 109–126.
- [20] Mewes, H. W., D. Frishman, U. B., Mannhaupt, G., Mayer, K. *et al.*, *Nucleic Acids Res.* 2002, 30, 31–34.
- [21] Wuchty, S., *Proteomics* 2002, 2, 1715–1723.
- [22] Wuchty, S., Oltvai, Z., Barabási, A.-L., *Nat. Genet.* 2003, 35, 176–179.
- [23] Hazbun, T., Fields, S., *Proc. Natl. Acad. Sci. USA* 2001, 98, 427–4278.
- [24] Hoffmann, R., Valencia, A., *Trends Genet.* 2003, 19, 681–683.
- [25] Albert, R., Jeong, H., Barabási, A., *Nature* 2000, 406, 378–382.