The yeast protein-protein interaction map is a highly modular network with a staircase community structure

Haijun Zhou and Reinhard Lipowsky
Max-Planck-Institute of Colloids and Interfaces,
14424 Potsdam, Germany

May 6, 2006

Abstract

Summary: The construction of genome wide protein–protein interaction maps makes it feasible to study the global organization of proteins in a biological cell. Here the module organization of the protein–protein interaction network (PPIN) of budding yeast are investigated by `Netwalk`, an algorithm based on biased random (Brownian) walks. The yeast PPIN is a highly modular network, it has a modularity value of $Q \simeq 0.62$. Measured by a dissimilarity index, the giant component of yeast PPIN contains 449 elementary modules that organize into a staircase community structure. The elementary modules at the lowest level of this staircase are mainly responsible for protein production; and an expanding central community is formed by the addition of other elementary modules at increasingly higher dissimilarity levels, which are responsible for protein synthesis regulation, protein transport, cell cycle control, and so on. Many proteins that function as bridges between different elementary modules are also identified.

Availability and supporting information: The `Netwalk` algorithm is available upon request (`zhou@mpikg-golm.mpg.de`). Information on the identified elementary modules is also deposited at http://www.mpikg-golm.mpg.de/theory/people/zhou/EModule.html.

During the last couple of years, genome wide protein–protein interaction maps of three model eukaryotic organisms, budding yeast *Saccharomyces cerevisiae* [Uetz *et al.*, 2000, Ito *et al.*, 2001, Gavin *et al.*, 2002, Ho *et al.*, 2002, Zhu *et al.*, 2001], fruit fly *Drosophila melanogaster* [Giot *et al.*, 2003] and worm *Caenorhabditis elegans* [Li *et al.*, 2004], have been determined. This rapid accumulation of protein–protein interaction information is largely due to recent advances in large–scale and high–throughput methods including yeast two–hybrid, protein mass spectrometry, and functional protein microarrays (for a review, see [Phizicky *et al.*, 2003]). In the case of budding yeast, the combination of high–throughput results and conventional small–scale observations leads to a large network of ∼16,000 protein–protein interactions involving ∼5,000 proteins [Xenarios *et al.*, 2002].

Complementary to this rapid progress, many theoretical efforts have been devoted to the yeast protein–protein interaction network (yeast PPIN) in the last few years. The first issue to be addressed was the reliability of protein interactions predicted by large–scale approaches. Different statistical methods were implemented to evaluate the accuracy of individual protein–protein interactions [Salwinski & Eisenberg, 2003, Mrowka *et al.*, 2001, von Mering *et al.*, 2002, Deane *et al.*, 2002, Goldberg & Roth, 2003], resulting in a highly confident and regularly updated CORE dataset for yeast [Deane *et al.*, 2002]. Another important issue is to predict protein functions based on yeast PPIN. Recently, redundancies and random false positives in the protein interactions of high–throughput experiments were exploited by network–based statistical methods, and functions of many unannotated proteins were predicted with high confidence [Samanta & Liang, 2003, Vazquez *et al.*, 2003, Karaoz *et al.*, 2004]. Furthermore, as far as *local* structures of yeast PPIN are concerned, many statistically significant protein modules were also predicted through network–based bioinformatics studies [Bu *et al.*, 2003, Rives & Galitski, 2003, Spirin & Mirny, 2003], where most of these modules correspond to protein complexes.

In this article, we address the *global* community structure and module organization of yeast PPIN. We find that yeast PPIN is a highly modular network; its modularity is 0.616, and 64.15% of the protein–protein interactions occur within different *elementary modules*. Using a quantitative dissimilarity measure, we show that the elementary modules of yeast PPIN form a staircase community structure. The elementary modules at the lowest steps of this staircase are mainly responsible for protein production; and an expanding central community is formed by the addition of other elementary

modules at increasingly higher dissimilarity levels, which are responsible for protein synthesis regulation, protein transport, cell cycle control, and so on. Many proteins that function as bridges between different elementary modules are also identified. We expect that the community structure of yeast PPIN reported in this paper will lead to a deeper understanding of budding yeast cell biology at the module level [Hartwell *et al.*, 1999].

The community structure of yeast PPIN was identified by `Netwalk`, an algorithm based on biased random walks on weighted networks [Zhou & Lipowsky, 2004]. `Netwalk` has the following properties: (i) Both the network's local and global (topological) structural properties are taken into account; (ii) The contributions of all paths are included, not just those of the geodesic ones as in [Girvan & Newman, 2002]; (iii) A quantitative dissimilarity measure is given for each pair of neighboring vertices or communities.

Methods

Dissimilarity measure and community structure

Here we briefly describe the algorithm `Netwalk` [Zhou & Lipowsky, 2004]. Consider a connected network of N vertices ($i = 1, \ldots, N$) and M edges. Each edge (i, j) has a positive weight ω_{ij}, which corresponds to the interaction strength between vertex i and vertex j. A Brownian particle moves on the network. At each step, it jumps from its present position, vertex i, to a neighboring vertex j with probability $P_{ij} \propto \omega_{ij}(c_{ij} + 1)^{\gamma}$, where c_{ij} is the number of common nearest neighbors of i and j, and $\gamma \geq 0$ is a bias coefficient. `Netwalk` achieves an optimal performance if one chooses the bias coefficient $\gamma = 1$ [Zhou & Lipowsky, 2004].

The mean first passage time d_{ij} is the average number of steps the Brownian particle takes to travel from vertex i to vertex j. Compared to the shortest–path distance as used in [Rives & Galitski, 2003], this mean first passage time integrates much more information on the network's architecture since contributions from all the non-geodesic paths between i and j are taken into account. The vector $\{d_{i1}, \ldots, d_{iN}\}$ can be regarded as the "coordinate" of vertex i in the network. For each edge (i, j), a dissimilarity index $\Lambda(i, j)$ is defined as [Zhou, 2003]

$$\Lambda(i, j) = \Big(\sum_{k \neq i,j} [d_{ik} - d_{jk}]^2 \Big)^{1/2} / (N - 2), \qquad (1)$$

which quantifies the extent of dissimilarity between vertices i and j (as demonstrated in this paper and in our previous work [Zhou, 2003, Zhou & Lipowsky, 2004], our dissimilarity measure in terms of Euclidean distance appear to work very well; other dissimilarity measures could also be explored). If i and j both belong to a group of vertices that are densely inter-connected (a community), their coordinates will be similar, and $\Lambda(i,j)$ will be small. The dissimilarity between two groups of vertices as provided, e.g., by two communities, is the mean value of $\Lambda(i,j)$ averaged over all edges (i,j) which provide direct links between them. $\Lambda(i,j)$ is computed exactly using matrix method [Zhou & Lipowsky, 2004].

Initially each vertex of the network is regarded as a community. At each step `Netwalk` merges the two communities that have the minimal dissimilarity index into a larger one, until all the vertices are merged together. The output is a dendrogram showing the relationship between different proteins and protein communities.

Identification of elementary modules

Following the idea of Newman and Girvan [Newman & Girvan, 2004], the modularity Q_α of a community α is defined as

$$Q_\alpha = \frac{I_\alpha}{M} - \frac{K_\alpha(K_\alpha - 1)}{2M(2M-1)}, \tag{2}$$

where I_α is the number of edges internal to community α and $K_\alpha = \sum_{i\in\alpha} k_i$ is its total vertex degree (k_i is the degree of vertex i, namely the number of edges incident to it). In Eq. (2), the first term after the equality is the actual fraction of edges internal to community α, and the second term is its expected value if the edges of the network are completely randomized while fixing the degree of each vertex.

During the community–merging process of `Netwalk`, the modularity $Q_{\alpha\cup\beta}$ of each merged community $\alpha\cup\beta$ is calculated first by Eq. (2). If the calculated value $Q_{\alpha\cup\beta} < Q_\alpha + Q_\beta$, then $Q_{\alpha\cup\beta}$ is assigned the new value $Q_{\alpha\cup\beta} = Q_\alpha + Q_\beta$; otherwise, $Q_{\alpha\cup\beta}$ is not changed and the merged community is marked as a "candidate". After the community dendrogram has been constructed, each branch of this dendrogram is backtracked from the top level, until a community with a "candidate" mark is encountered. This community, which may contain other 'candidates' as smaller subsets, is then considered to represent an *elementary module* of the network. As a result of this backtrack process,

the network is partitioned into a set of mutually exclusive elementary modules $\alpha_1, \alpha_2, \ldots, \alpha_l$ such that the total modularity $Q = \sum_{\alpha_i} Q_{\alpha_i}$ achieves its *global* maximum value. This elementary module identification procedure can be easily integrated into `Netwalk`. We emphasize that, the elementary modules determined in this way do not correspond to a partition of the network at a fixed level of the dissimilarity index.

`Netwalk` has been tested on some artificial and real–world modular networks [Zhou & Lipowsky, 2004, Zhou, 2003]. To test its performance on networks with long tailed degree distributions such as yeast PPIN, we have further applied `Netwalk` on an ensemble of growing modular networks. An artificial network is generated by adding a small module of 6 vertices and $10-11$ internal edges to the existing network at each growth step. Using preferential attachment rule [Barabási & Albert, 1999], $5-6$ edges are set up between the newly added module and the old network. Each random network has the same number of vertices and edges as in yeast PPIN and it also has a long tailed vertex degree profile by construction [Barabási & Albert, 1999]. A total number of 100 random networks were generated. When `Netwalk` is applied to such a network, it predicts a certain number of elementary modules. If two vertices belong to the same artificially constructed module, they also belong to the same predicted elementary module with probability 0.95 ± 0.01 (averaged over 100 network realizations); in addition, $85\% \pm 2\%$ of the predicted elementary modules are identical to one of the constructed artificial modules.

Stability of an elementary module

After all the elementary modules are reported by *Netwalk*, each of them is assigned two stability parameters. The first stability measure is provided by the affinity A_α of an elementary module α as defined by

$$A_\alpha = (\sum_{i \in \alpha} k_i^{(\alpha)}/k_i)/|\alpha|, \quad \text{with } 0 \leq A_\alpha \leq 1 \tag{3}$$

where $k_i^{(\alpha)}$ is the number of vertices in module α that interact directly with vertex i, and $|\alpha|$ is the size of module α. The affinity A_α measures how strongly a vertex of module α interacts with other vertices in α compared with its interaction with vertices outside α.

As the second stability measure, we follow Hopcroft and co-authors [Hopcroft *et al.*, 2004] and ask to what extent vertices in each elementary module tend to be together

under perturbations of the network architecture. As a type-I perturbation, a randomly selected 5% of the total edges are removed; as a type-II perturbation, the weights of all the edges are assigned random values according to some prescribed distribution. We then use `Netwalk` to find all the elementary modules of the perturbed network. The integrity L_α of an elementary module α of the original network is defined as

$$L_\alpha = \max_{\alpha'}(|\alpha \cap \alpha'|/\sqrt{|\alpha||\alpha'|}), \quad \text{with } 0 \leq L_\alpha \leq 1 \tag{4}$$

where α' is an elementary module of the perturbed network, and $\alpha \cap \alpha'$ means the intersection of module α and α'. The integrity L_α indicates the stability of module α against network perturbations. In this paper, 100 samples of perturbed networks of each type were generated to estimate the integrity values for each elementary module in yeast PPIN.

Results

The yeast protein–protein interaction map (`Scerecr20040104.tab`) was extracted from the Database of Interacting Proteins DIP (`http://dip.doe-mbi.ucla.edu`) [Xenarios *et al.*, 2002, Deane *et al.*, 2002]. The giant component of this network has $N = 2406$ proteins (vertices) and $M = 6117$ protein–protein interactions (edges), with an average clustering coefficient of 0.3 [Watts & Strogatz, 1998]. Because of the lack of detailed information on the strengths of protein–protein interactions, each edge of the network is assigned a uniform weight $\omega = 1$ in this study. The `Netwalk` algorithm is then applied to this giant component to construct its community structure (see *Methods*) as shown in Fig. 1 and Fig. 2. The CPU running time for `Netwalk` is about five minutes.

Based on this community structure, the proteins are then partitioned into a set of *elementary modules* in such a way that the modularity of the network attains its global maximum value (see *Methods*). This partition consists of 449 elementary modules (Table 1; a complete list is shown in *Supporting information*). The modularity is $Q = 0.616$, and 64.15% of the edges are internal edges (both of the two incident vertices of such an edge belong to the same elementary module). Furthermore, we found that the modularity of yeast PPIN approaches to the value 0.659 ± 0.001 (averaged over 100 realizations) of a growing modular network with the same number of vertices and edges (see *Methods*). Based on this fact and the empirical observation

by Newman and Girvan [Newman & Girvan, 2004] that real–world modular networks usually have modularity values $0.3 < Q < 0.7$, we conclude that yeast PPIN is a highly modular network. This finding strongly supports and substantiates the conjecture of Hartwell *et al.* [Hartwell *et al.*, 1999] that cell biology is modular. We noticed that many of the statistically significant modules identified by other authors [Bu *et al.*, 2003, Rives & Galitski, 2003, Spirin & Mirny, 2003], such as the nuclear pore complex, the RNA polymerase holoenzymes, the RNA exosome, the anaphase–promoting complex, etc., are subgroups of the elementary modules predicted by `Netwalk` (see *Supporting information*). This gives further confirmation of the biological significance of the elementary modules determined here. The elementary modules are also very stable in terms of affinity and integrity measures (see *Methods* and *Supporting information*), except for those containing less than three proteins.

The organization of the 449 elementary modules of yeast PPIN is shown in the inset of Fig. 1. It shows a clear staircase structure. The formation of yeast PPIN can therefore be described by the following way: (a) The 18 elementary modules `M001-M018` form a merged community with internal dissimilarity indices all less than $\Lambda = 1.256$. (b) The 24 elementary modules `M019-M042` form another merged community with internal dissimilarity indices all less than $\Lambda = 1.262$. (c) These two merged communities then merge into a community of even larger size at dissimilarity value $\Lambda = 1.288$. (d) Then the elementary module `M043` merges with the above–mentioned community to form a larger community at dissimilarity value $\Lambda = 1.330$. This is followed by the addition of `M044` at $\Lambda = 1.347$, `M045` at $\Lambda = 1.391$, ..., and `M449` at $\Lambda = 31.992$. The biological interpretation of this staircase organization will be discussed in the next section.

To exclude the possibility that the high modularity of yeast PPIN arises from a long–tailed vertex degree distribution [Barabási & Albert, 1999], we follow Maslov and Sneppen [Maslov & Sneppen, 2002] and generate an ensemble of randomized networks. All of these networks have completely random edge connections but have the same degree sequence as the real network and are connected. These randomized networks have a very low modularity of 0.259 ± 0.004 (over 100 realizations). Therefore, the high modularity of yeast PPIN is a consequence of strong correlation among the protein–protein interactions, which, in turn, has only been achieved through the long history of biological evolution and selection.

To exclude the possibility that the community structure of yeast PPIN is

an artifact caused by the uniform edge weight distribution used in this work, we have studied the community structure of a modified yeast PPIN, in which each edge can have the weight $\omega = 0.5$ and 1.5 with equal probability. We found that, as quantified by our integrity measure, the community structure of the modified network is essentially the same as that for the network with $\omega \equiv 1$ (*Supporting information*). Furthermore, we found that the modularity value of the modified network has a slightly changed value of 0.618 ± 0.005 and the fraction of internal edges is now 0.651 ± 0.008 (averaged over 100 realizations). The `Netwalk` algorithm is insensitive to edge weight perturbations, because these perturbations only affect the local property of the network. Some knowledge about the reliability of protein–protein interactions is also documented in the DIP database. This information can be incorporated in the edge weight to improve the performance of the algorithm.

Biological interpretation

Of the 449 predicted elementary modules, 46 contain ten or more proteins. The positions of these elementary modules in the community structure are highlighted with colors in Fig. 1, and the direct interaction patterns among these elementary modules is shown in Fig. 3.

To understand more about the module organization of yeast PPIN, hereafter we look at the biological functions of those elementary modules with fifteen or more proteins in more detail (Table 1). Information of each protein's function and involved biological processes is extracted from the Comprehensive Yeast Genome Database [Mewes *et al.*, 2002].

The major function of `M001` is nuclear transport of protein and RNA. The translation initiation factor `eIF3` also belongs to it. `M003` contains the 26S proteasome. Besides its role in protein degradation, the importance of the proteasome to transcription elongation and termination is well documented [Gillette *et al.*, 2004]. `M016` in involved in pre-mRNA splicing/processing. `M019` is involved in rRNA splicing and ribosome biogenesis. `M020` contains many transcription factors, including the Mediator complex and the TFIID/SAGA complex.

These five elementary modules therefore include groups of proteins for transport between cell nucleus and cytoplasm, RNA transcription control, RNA splicing, ribosome biogenesis, protein synthesis and protein degradation. They interact strongly with each other and with other elementary mod-

ules (Fig. 3). They belong to the lowest steps in the staircase community structure. The fraction of lethal (essential) genes corresponding to proteins in each of these modules is larger than 30.4%, which corresponds to the fraction of lethal genes for the whole network [Giaever *et al.*, 2002]. Especially, 73.8% of the proteins in the largest elementary module `M019` are essential. Taking together, the lowest step of the community structure of yeast PPIN represents the fundamental protein production (translation) "factory" of budding yeast. It can be regarded as the *core* of yeast PPIN, since (1) it is located at the deepest level of the staircase community structure; and (2) other elementary modules will associate with it in a consecutive order to form a more and more larger community.

The functions of other elementary modules outside the above–mentioned protein production core are more divergent. The main function categories of these *peripheral* elementary modules are:

mRNA decay and modification. Examples include `M084`, `M085`, and `M133`.

cell cycle control. Examples include `M052`, `M088`, `M136`, `M174`, `M182`, `M205`, `M233`, and `M356`.

cytoplasmic protein transport. Examples include `M056`, `M076`, `M133`, `M154`, `M381`, and `M425`.

DNA processing. Examples include `M168`, `M265`, and `M268`.

ATP synthesis. This includes `M448`.

Therefore, on the global scale the protein–protein interaction network of yeast can be described as consisting of a protein production core and a peripheral part that, among other functions, regulate the protein synthesis process and transport protein to different destinations. This division, however, should not be regarded as very strict. For example, we notice that the elementary modules `M103` (containing RNA polymerases) and `M244` (containing translation initiation factor eIF2) are located outside the core part of protein production.

In what follows we mention some further aspects of the community structure of yeast PPIN.

The elementary module `M052` includes Cdc42, Cdc20 and many other proteins responsible for the pheromone pathway. According to the community

structure Fig. 1, M052 is located adjacent to the above–mentioned protein production core. This arrangement may enable the cell to change its morphology rapidly in response to even a weak or transient sex-attractant signal. As is consistent with this requirement, the pheromone pathway strongly connect to the septin complex (M073) and to the actin–associated cytoskeleton (M076) (Fig. 3).

The cell cycle of budding yeast is controlled by many elementary modules, the most important example being M088, the Cdc28–dependent pathway. M088 is responsible for the G1/S and G2/M transition; it has relatively low integrity and affinity values in comparison with other elementary modules (see *Supporting information*), possibly reflecting the fact that many protein–protein interactions in this module do not occur at the same time and location [Spirin & Mirny, 2003, Alberts *et al.*, 1989]. M088 interacts strongly with M052, which is also responsible for cell budding.

DNA replication, chromosome segregation, and chromatin assembly are accomplished by M168 (DNA repair), M265 (DNA replication), M268 (chromosome segregation), M269 (origin recognition complex), M344 (anaphase–promoting complex), M356 (spindle pole body), and M358 (chromatin assembly). These modules are located far away from the protein production core (Fig. 1), consistent with the finding that protein synthesis is severely inactivated during chromosome segregation [Alberts *et al.*, 1989].

The most remote modules include M366 (protein import into ER), M381 (peroxin), M425 (mitochondrial protein import), and M448 (mitochondrial ATP synthesis). They are needed for protein import into cellular organelles and for mitochondrial ATP synthesis.

Among the 449 elementary modules of the yeast PPIN, 119 contains just a single protein (*Supporting Information*). The vertex degrees of these 119 proteins range from 2 to 18, with a mean value 4.72, and on average 87% of the edges of such a protein are connected to different elementary modules. For example, Clu1 (M009, vertex degree 13) is connected to 12 modules; Tif2 (M012, 18) is connected to 14 modules (both Clu1 and Tif2 are translation initiation factor eIF3 subunits [Mewes *et al.*, 2002]); Cct5 (M066, 15) is connected to 13 modules; Rvb2 (M114, 16) is connected to 13 modules; and Mis1 (M139, 11) is connected to 11 modules. It is very likely that such proteins play an important role in mediating the interactions between different elementary modules. There are also 137 elementary modules of size two. These "modules" may also serve as bridges between different elementary modules.

Conclusion and outlook

In this paper we have studied the community structure of the protein–protein interaction network of budding yeast with `Netwalk`. We found that this network is highly modular, with a modularity of 0.616. The elementary modules of the network are organized into a staircase community structure. At the deepest level of this staircase are those elementary modules that are directly responsible for protein synthesis. These core modules are surrounded by many peripheral modules that regulate protein production in response to internal or external signals and transport protein products to different parts of the cell. There exist also many proteins which interact with different elementary modules and serve as bridges between them.

The community structures of the protein–protein interaction networks of fruit fly [Giot *et al.*, 2003] and worm [Li *et al.*, 2004] were also studied by us and found to be similar to Fig. 1 of budding yeast. Similar organizations have been inferred for many metabolic networks [Ravasz *et al.*, 2002, Holme *et al.*, 2003] and for an integrated genetic network of budding yeast [Tanay *et al.*, 2004]. The full biological significance of such a structure is yet to be more thoroughly interpreted. We notice here that the global organization of the budding yeast protein–protein interaction network into a protein production part (core) and a product regulation/transport part (periphery) is similar to the organization of many economic enterprises. This global organization of a biological cell's protein–protein interaction network may reflect the same optimization principles as that of an economic system: (i) to proliferate at maximal rate in appropriate conditions (quick in response, or sensitivity); (ii) be able to survive in severe conditions (robustness). The position of each elementary module in Fig. 1 may also contain some evolutionary information. It is interesting to check whether elementary modules at the more left part of Fig. 1 are generally more "old" or not.

When the yeast PPIN is viewed at the elementary module level, it forms a network such as Fig. 3 with edges between these modules. An analysis of the functional properties of these cluster–linking edges may also be very valuable.

In the work presented here, the protein–protein interaction network was analyzed as a static structure. In the future as more biological data are accumulated, information on the dynamics of the protein–protein interactions will be available. This information can be incorporated into the network model in order to address the dynamics of the module organization. Also

the protein–protein interaction strength or reliability can be incorporated in a properly defined weight matrix. A further extension of the present work may be to calculate the dissimilarity index $\Lambda(i, j)$ between any two arbitrarily chosen proteins i and j. This will enable us to cluster the yeast PPIN with complete–linkage methods.

References

[Alberts *et al.*, 1989] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1989) *Molecular Biology of the Cell.* 2nd edition,, Garland publishing, New York.

[Barabási & Albert, 1999] Barabási, A.-L. & Albert, R. (1999) Emergence of scaling in random networks. *Science,* **286**, 509–512.

[Bu *et al.*, 2003] Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G. & Chen, R. (2003) Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res.,* **31**, 2443–2450.

[Deane *et al.*, 2002] Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high-throughput observations. *Mol. Cell. Proteomics,* **1**, 349–356.

[Gavin *et al.*, 2002] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* **415**, 141–147.

[Giaever *et al.*, 2002] Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A. P. *et al.* (2002) Functional profiling of the *saccharomyces cerevisiae* genome. *Nature,* **418**, 387–391.

[Gillette *et al.*, 2004] Gillette, T. G., Gonzalez, F., Delahodde, A., Johnston, S. A. & Kodadek, T. (2004) Physical and functional association of rna polymerase ii and the proteasome. *Proc. Natl. Acad. Sci. USA,* **101**, 5904–5909.

[Giot *et al.*, 2003] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P. *et al.* (2003) A protein interaction map of *drosophila melanogaster*. *Science,* **302**, 1727–1736.

[Girvan & Newman, 2002] Girvan, M. & Newman, M. E. J. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.,* **99**, 7821–7826.

[Goldberg & Roth, 2003] Goldberg, D. S. & Roth, F. P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA,* **100**, 4372–4376.

[Hartwell *et al.*, 1999] Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999) From molecular to modular cell biology. *Nature,* **402**, C47–C52.

[Ho *et al.*, 2002] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C. *et al.* (2002) Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature,* **415**, 180–183.

[Holme *et al.*, 2003] Holme, P., Huss, M. & Jeong, H. (2003) Subnetwork hierarchies of biochemical pathways. *Bioinformatics,* **19**, 532–538.

[Hopcroft *et al.*, 2004] Hopcroft, J., Khan, O., Kulis, B. & Selman, B. (2004) Tracking evolving communities in large linked networks. *Proc. Natl. Acad. Sci. USA,* **101**, pnas.0307750100.

[Ito *et al.*, 2001] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA,* **98**, 4569–4574.

[Karaoz *et al.*, 2004] Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R. & Kasif, S. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl. Acad. Sci. USA,* **101**, 2888–2893.

[Li *et al.*, 2004] Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S. *et al.* (2004) A map of the interactome network of the metazoan *c. elegans. Science,* **303**, 540–547.

[Maslov & Sneppen, 2002] Maslov, S. & Sneppen, K. (2002) Specificity and stability in topology of protein networks. *Science,* **296**, 910–913.

[Mewes *et al.*, 2002] Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkoetter, M., Rudd, S. & Weil, B. (2002) Mips: a database for genomes and protein sequences. *Nucleic Acids Res.,* **30**, 31–34.

[Mrowka *et al.*, 2001] Mrowka, R., Patzak, A. & Herzel, H. (2001) Is there a bias in proteome research? *Genome Res.,* **11**, 1971–1973.

[Newman & Girvan, 2004] Newman, M. E. J. & Girvan, M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E,* **69**, 026113.

[Phizicky *et al.*, 2003] Phizicky, E., Bastiaens, P. I. H., Zhu, H., Snyder, M. & Fields, S. (2003) Protein analysis on a proteomic scale. *Nature,* **422**, 208–215.

[Ravasz *et al.*, 2002] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science,* **297**, 1551–1555.

[Rives & Galitski, 2003] Rives, A. W. & Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA,* **100**, 1128–1133.

[Salwinski & Eisenberg, 2003] Salwinski, L. & Eisenberg, D. (2003) Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.,* **13**, 377–382.

[Samanta & Liang, 2003] Samanta, M. P. & Liang, S. (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. USA,* **100**, 12579–12583.

[Spirin & Mirny, 2003] Spirin, V. & Mirny, L. A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA,* **100**, 12123–12128.

[Tanay *et al.*, 2004] Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. USA,* **101**, 2981–2986.

[Uetz *et al.*, 2000] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature,* **403**, 623–627.

[Vazquez *et al.*, 2003] Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003) Global protein function prediction from protein-protein interaction networks. *Nature Biotech.,* **21**, 697–700.

[von Mering *et al.*, 2002] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature,* **417**, 399–403.

[Watts & Strogatz, 1998] Watts, D. J. & Strogatz, S. H. (1998) Collective dynamics of 'small-world' network. *Nature,* **393**, 440–442.

[Xenarios *et al.*, 2002] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kum, S. M. & Eisenberg, D. (2002) Dip: the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.,* **30**, 303–305.

[Zhou, 2003] Zhou, H. (2003) Distance, dissimilarity index, and network community structure. *Phys. Rev. E,* **67**, 061901.

[Zhou & Lipowsky, 2004] Zhou, H. & Lipowsky, R. (2004) Network brownian motion: a new method to measure vertex-vertex proximity and to identify communities and subcommunities. *Lect. Notes Comput. Sci.,* **3038**, 1062–1069.

[Zhu *et al.*, 2001] Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P. *et al.* (2001) Global analysis of protein activities using proteome chips. *Science,* **293**, 2101–2105.

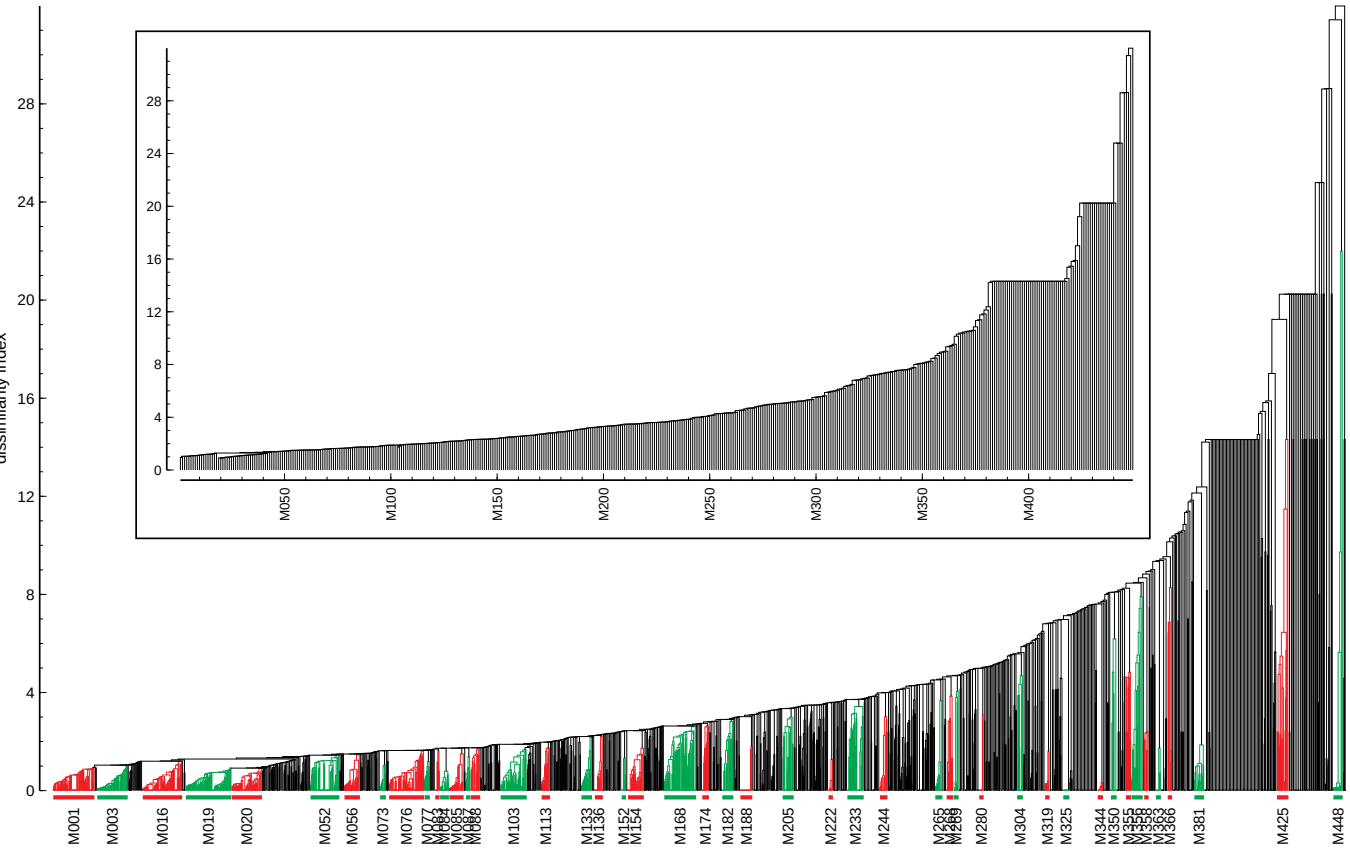

Figure 1: Community structure of the yeast protein-protein interaction network as predicted by the algorithm `Netwalk`. The positions of those elementary modules of size ≥ 10 are highlighted alternatively with red and green colors and their indices are listed. (Inset) staircase organization of the 449 elementary modules.

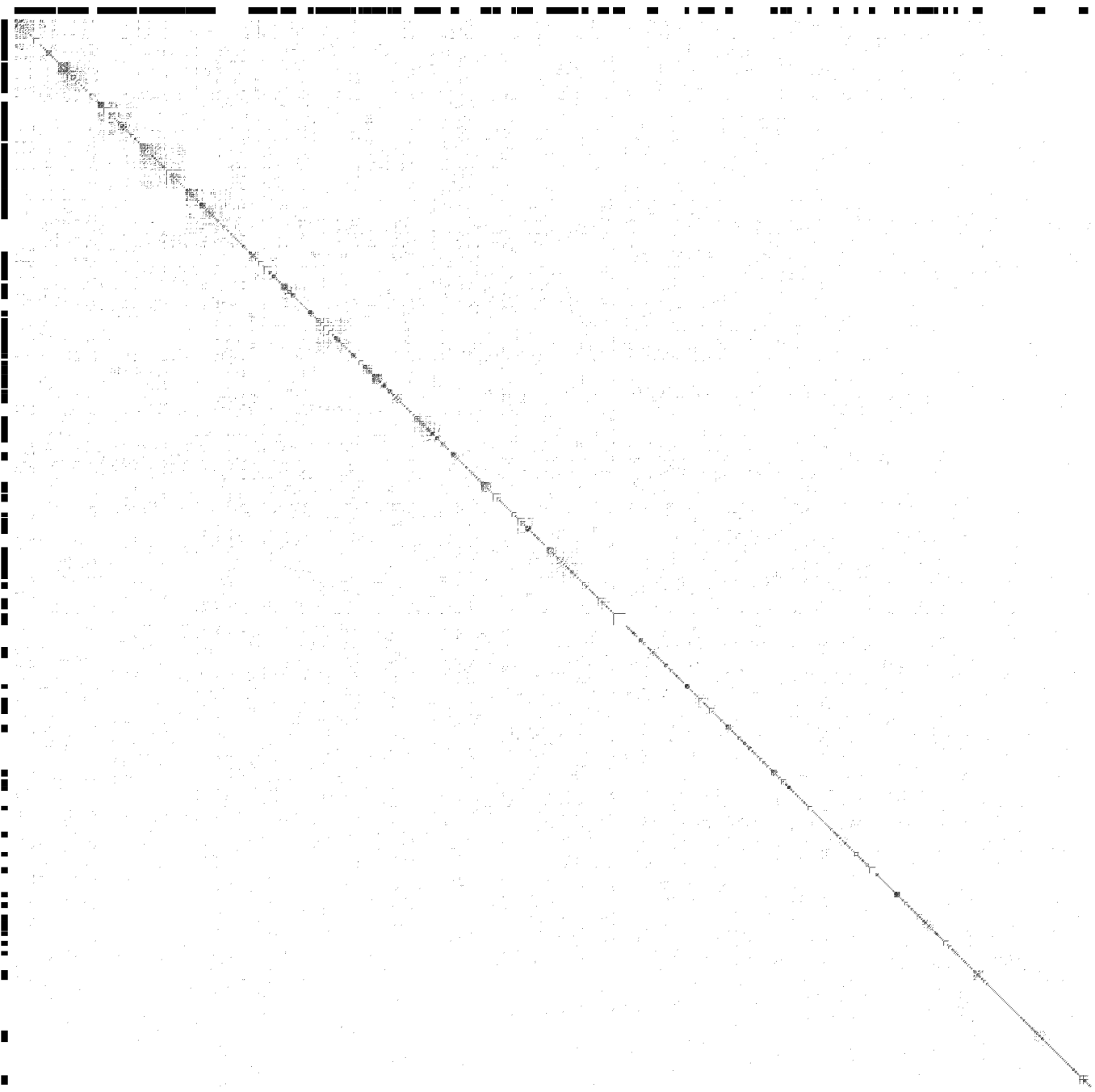

Figure 2: Two–dimensional illustration of protein–protein interaction (tiny dots). The protein of each protein is ordered according to the community structure Fig. 1. The horizontal and vertical bars mark the positions of those elementary modules of size ≥ 10 (the order is `M001`,..., `M448` left–right and top–down).

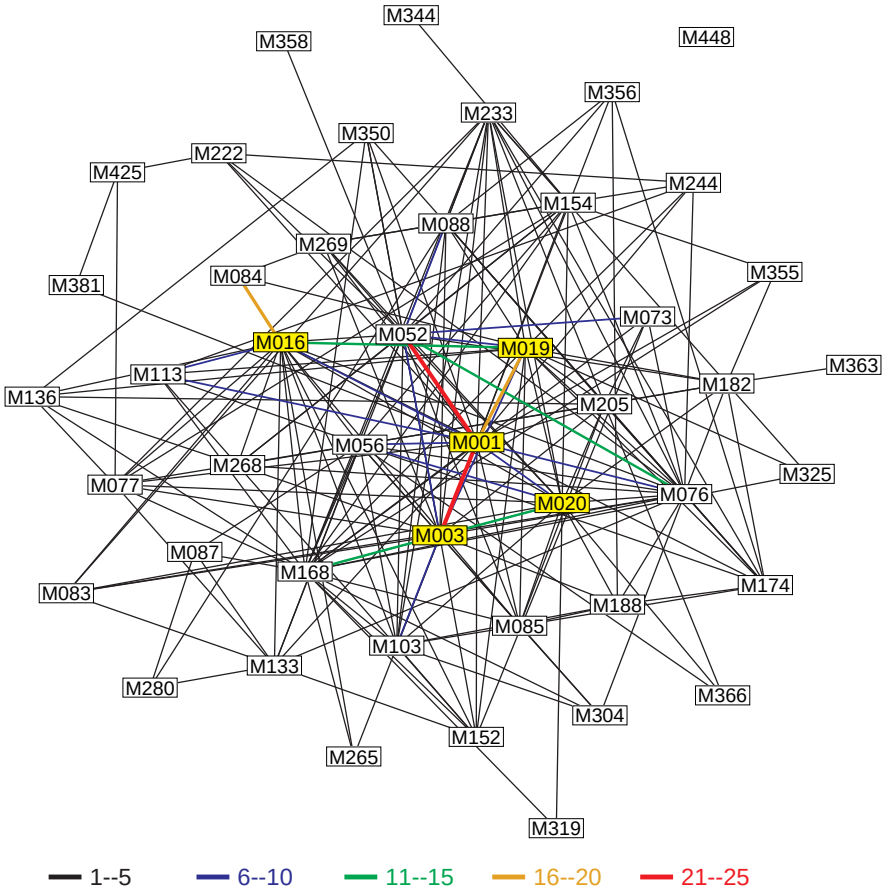

Figure 3: Connection pattern for the subset of elementary modules of the yeast protein-interaction network with size ≥ 10. An edge is drawn between two elementary modules if at least one protein-protein interaction exists between them, and the total number of such interactions between two elementary modules is indicated by a color code as indicated in the figure. Elementary module `M448` (ATP synthase) is isolated, i.e., all interactions between `M448` and the other elementary modules of this subset are mediated by other smaller elementary modules not included in this subset. The five yellow modules are members of the core of the network.

Table 1: Biological functions of each elementary module that contains fifteen or more proteins. The fraction of lethal genes [Giaever *et al.*, 2002] corresponding to the proteins of each module is also listed. In the fourth column the indices of the three most frequently occurred (groups of) major functional categories together with their occurring times are listed for each elementary module. More detailed information on the protein functional annotations as obtained from [Mewes *et al.*, 2002] is presented in *Supporting information*. The correspondence between each functional category and its index is as follows: **01**: metabolism; **02**: energy; **10**: cell cycle and DNA processing; **11**: transcription; **12**: protein synthesis; **14**: protein fate (folding, modification, destination); **20**: cellular transport, transport facilitation and transport routes; **32**: cell rescue, defense and virulence; **34**: interaction with cellular environment; **40**: cell fate; **42**: biogenesis of cellular components; **43**: cell type differentiation; **98**: classification not yet clear-cut; **99**: unclassified proteins.

index	size	lethal (%)	functional categories	most prominent function
M001	93	42.0	45(**20**), 16(**11**), 15(**12**)	nuclear transport
M003	69	46.4	35(**14**), 16(**10**), 10(**99**)	cytoplasmic and nuclear protein degradation
M016	89	46.1	56(**11**), 11(**01**), 8(**99**)	mRNA splicing
M019	103	73.8	38(**11**), 37(**98/99**), 12(**12**)	rRNA processing
M020	68	32.4	58(**11**), 15(**01**), 7(**10**)	transcriptional control
M052	65	9.2	25(**42**), 22(**40**,**43**), 15(**10**)	cell growth/morphogenesis; cytoskeleton; budding, cell polarity and filament formation
M056	35	57.1	27(**20**), 4(**14**,**99**), 2(**11**)	vesicular transport (Golgi network, etc.)
M076	79	13.9	34(**43**), 28(**42**), 27(**40**)	budding, cell polarity and filament formation
M084	20	60.0	13(**11**), 4(**10**,**98/99**), 1(**32**,**42**)	mRNA splicing
M085	32	50.0	18(**11**), 10(**99**), 2(**10**,**14**,**20**)	mRNA processing (splicing, 5'-, 3'-end proces
M088	21	14.3	10(**10**), 6(**34**,**43**), 5(**40**)	mitotic cell cycle and cell cycle control
M103	59	52.5	45(**11**), 6(**10**), 5(**14**)	rRNA synthesis; general transcription activiti
M113	19	42.1	8(**11**), 5(**01**), 3(**10**,**98/99**)	rRNA processing
M133	24	29.2	18(**20**), 5(**14**,**98/99**), 3(**42**,**43**)	vesicular transport (Golgi network, etc.)
M136	18	0.0	13(**11**), 4(**43**,**99**), 3(**10**,**34**)	transcriptional control
M154	36	19.4	12(**20**), 11(**14**), 10(**34**)	cation transport (Na, K, Ca , NH4, etc.); transport ATPases; vacuolar transport
M168	72	22.2	39(**10**), 18(**11**), 8(**42**)	DNA recombination and DNA repair
M174	15	20.0	7(**11**), 5(**10**), 4(**01**)	transcriptional control
M182	25	4.0	11(**11**), 9(**01**), 4(**99**)	transcriptional control; regulation of C-compound and carbohydrate utilization
M188	27	14.8	9(**01**), 5(**20**), 4(**10**,**34**,**40**,**99**)	
M205	25	16.0	12(**01**), 10(**10**,**34**), 9(**11**)	transcriptional control
M233	37	10.8	12(**10**), 10(**01**), 7(**11**)	mitotic cell cycle and cell cycle control
M244	17	52.9	8(**12**), 4(**20**), 3(**01**)	translation
M265	16	37.5	11(**10**), 2(**99**), 1(**01**,**14**,**98**)	DNA synthesis and replication
M268	15	53.3	9(**10**), 3(**14**,**42**), 2(**02**)	chromosome segregation/division
M356	25	56.0	17(**10**), 9(**42**), 5(**14**)	mitotic cell cycle and cell cycle control
M381	22	0.0	12(**20**), 9(**14**), 4(**01**,**98/99**)	peroxisomal transport
M425	25	53.8	18(**14**), 17(**20**), 6(**42**)	mitochondrial transport
M448	21	4.8	17(**02**), 14(**20**), 12(**34**)	respiration