Supplementary Note

For the article "***Evolutionary conservation of motif constituents within the yeast protein interaction network***" by S. Wuchty, Z.N. Oltvai and A.-L. Barabási.

Table of content

A. Databases

Protein Interaction Databases: Large-scale two-hybrid screens, which allow the identification of potential protein-protein interactions between open reading frames predicted from the *S. cerevisiae* genomic sequence[1,2] are an integral part of proteomics research. Yet, the quality of two-hybrid data is significantly affected by high rates of false positives and false negatives[3], indicated also by the fact that the results obtained by different groups overlap only to a limited extent[4]. Moreover, many identified interactions rely on positive signals from a single technique and result from indirect observations.

A variety of databases provides efficient and easy access to the rapidly increasing knowledge about various protein interaction processes. The MIPS database[5] collects genetic, biochemical and cell biological knowledge of various organisms which was extracted from the literature. BioKnowledge library is a composition of protein-specific information collected from the scientific literature[6]. Schwikowski et al. gathered all available protein interaction data of yeast, providing a comprehensive graph of protein-protein interactions[7].

In our study, we used the database of interacting proteins (DIP) which is based on extensive literature searches, and aims to provide the best curated collection of all functional linkages of proteins obtained by experimental methods. The majority of protein-protein interaction data relies on yeast two-hybrid- and co-immunoprecipitation experiments. 84% of the interactions are detected by only one single experiment, whereas 16% are confirmed by more than one experimental method. DIP records nearly 3000 proteins which are involved in approximately 9000 interactions[8]. Since it is manually curated, DIP provides high quality interaction data by minimizing the total number of false positive and negative interactions.

Orthologous Sequence Information: Methods of finding orthologous pairs of sequences often utilize pairwise BLAST comparisons of whole proteomes. Each protein represents

queries against the entire proteome of the other species. Symmetrical best hits in these BLAST searches, emphasizing expectation values smaller than 10^{-3}, were considered to be orthologous. The database of clusters of orthologous genes (COG) was compiled[9] utilizing such orthologous sequence pairs of mainly prokaryotic organisms.

Our choice of orthologous protein sequence information is the InParanoid database[10] which, similarly to COG, runs an all-versus-all BLAST search with two sets of sequences. Sequence pairs with mutual best hits are detected and serve as central main ortholog pairs around which further orthologs from both species are clustered in later steps. The initial assumption is that sequences from the same species that are more similar to the main ortholog than to any sequence from other species are 'in-paralogs', belonging to the same group of orthologs. In contrast to COG, the quality of the resulting orthologous clusters is examined and increased by a final bootstrap analysis[10]. Furthermore, InParanoid provides comprehensive pairwise comparative orthologous information between *S. cerevisiae* and *H. sapiens, D. melanogaster, C. elegans, M. musculus* and *A. thaliana*, which are absent in COG.

B. Effects of data incompleteness and errors

B.1 Errors in protein-protein interactions (Supplementary Table 1)

It is well known that protein-protein interaction and ortholog databases are incomplete and contain false positives[3]. In order to address the potential effects of this technology-derived noise we investigated the effects of data incompleteness and incorrectness on our results. We mimick false positives by the addition of extra 10% or 20% of interactions between randomly picked protein pairs that were previously absent in the yeast protein interaction network. In turn, the random removal of 10% and 20% of known interactions accounts for data incompleteness (false negatives). Subsequently, we recalculated the conservation rate of the motifs for the artificially altered database (Supplementary Table 1). The measurements indicate that the conservation rate of the motifs is only marginally affected, and the trend towards the increased degree of conservation of larger motifs remains unchanged. In particular, while data incompleteness (false negatives) has negligible influence, since false positives randomize the network, the motif conservation is slightly, -but not significantly-, affected.

B.2 Errors in ortholog protein assignment (Supplementary Table 2)

Similar results are obtained for the introduction of false positive and negative orthologs. For this we removed 10% or 20% of the assigned ortholog proteins in order to simulate the effects of false negative ortholog data. Accounting for false positives, we randomly added 10% or 20% more proteins to the original set of orthologous proteins. The results of the subsequent recalculation of the motif's conservation rate are compiled in Supplementary Table 2, indicating that the addition of false positive orthologs leaves the conservation rates almost unaltered. In contrast, false negatives lead to decreased conservation rates, since the removal of known orthologs significantly lowers the statistics. However, the basic trend observed in

the paper remains valid: Despite the systematic lowering in the conservation rate for false negatives, the highly connected motifs are more likely conserved than their less cohesive counterparts. Although flawed interaction and orthologs data obviously affect the exact quantities, we conclude that qualitative features of the networks, such as trends towards evolutionary conservation of the motifs remain robust.

B3: The effect of protein complexes (Supplementary Table 3):

Protein complexes, in which each protein interacts with almost all other complex partners, might bias our investigation and could be viewed as the sole origin of the observed conservation of the highly connected motifs. In the following, we show that the presence of protein complexes cannot explain exclusively the observed conservation effects. To eliminate the effects of protein complexes we removed all proteins that are part of known complexes based on the Swissprot database[11] (31 proteins and 564 interactions) and recalculated the conservation rates of motifs remaining in the diluted network. Although we observe, as expected, a common decrease of conservation rates, the basic trend towards higher conservation rates of strongly connected motifs remains unaltered. In comparison to the random reference sets, natural conservation ratios of the real motifs are still several orders of magnitude higher.

		False Negatives - 20 % interactions	False Negatives - 10 % interactions	Natural	False Positives + 10 % interactions	False Positives + 20 % interactions
#	Motifs	Conservation rate	Conservation rate	Conservation rate	Conservation rate	Conservation rate
1		13.46 %	13.46 %	13.67 %	12.83 %	12.23 %
2		5.02 %	5.27 %	4.99 %	4.81 %	4.71 %
3		21.66 %	21.92 %	20.51 %	16.83 %	15.28 %
4		0.82 %	0.71 %	0.73 %	0.76 %	0.73 %
5		2.83 %	2.67 %	2.64 %	2.50 %	2.38 %
6		7.53 %	10.73 %	6.71 %	6.58 %	6.46 %
7		8.82 %	7.52 %	7.67 %	6.83 %	6.26 %
8		21.64 %	19.14 %	18.68 %	18.46 %	18.46 %
9		39.81 %	24.40 %	32.53 %	26.63 %	23.00 %
10		16.20 %	21.81 %	14.77 %	14.64 %	14.57 %
11		43.98 %	49.03 %	47.24 %	37.14 %	24.21 %

Supplementary Table 1: Robustness of motif constituents upon incomplete or erroneous protein interaction data. As in Supplementary Table 1, the third column denotes the number of motifs found in the interaction network of 3,128 yeast proteins, obtained by counting subgraphs of two to five nodes. Our set of orthologs embraces again all 678 proteins that have an orthologue in each proteome of human, mouse, fly, worm and *A. thaliana*. The natural conservation rate shows what fraction of the original yeast motifs are evolutionary fully conserved. In order to investigate the influence of partially flawed interaction data on the conservation rate of motifs, 10% of the underlying interaction network's edges were randomly chosen and deleted mimicking false negative interaction signals. Conservation rates of motifs were subsequently calculated (col. 5). The same procedure was repeated after randomly deleting 20% of the edges of the initial protein interaction network (col. 4). Accounting for false positive signals, 10% (col. 6) and 20% (col. 7) more new edges, previously absent from the interaction network were randomly added and conservation rates of motifs thus obtained. Although the numbers of newly added or removed edges are quite high, the overall trend that highly intraconnected motifs persist is not affected. Although it is commonly assumed that yeast protein interaction data are highly incomplete and flawed due to the shortcomings of the widely used two-hybrid experimental set up, conservation rates, in particular for the triangles #3, squares #9 and pentagons #11, appear to be quite robust.

		False Negatives - 20 % orthologs	False Negatives - 10 % orthologs	Natural	False Positives + 10 % orthologs	False Positives + 20 % orthologs
#	Motifs	Conservation rate	Conservation rate	Conservation rate	Conservation rate	Conservation rate
1		9.03 %	11.00 %	13.67 %	14.60 %	15.95 %
2		2.55 %	3.83 %	4.99 %	5.49 %	6.15 %
3		9.93 %	17.13 %	20.51 %	21.71 %	21.48 %
4		0.23 %	0.51 %	0.73 %	0.83 %	1.40 %
5		0.90 %	1.81 %	2.64 %	2.89 %	4.04 %
6		1.99 %	4.04 %	6.71 %	6.89 %	8.23 %
7		2.00%	5.00 %	7.67 %	8.10 %	10.12 %
8		3.37 %	11.00 %	18.68 %	18.96 %	21.81 %
9		4.57 %	17.79 %	32.53 %	32.58 %	36.30 %
10		7.42 %	7.57 %	14.77 %	15.72 %	14.93 %
11		11.72 %	16.82 %	47.24 %	49.13 %	47.24 %

Supplementary Table 2: Robustness of motif constituents upon incomplete or erroneous orthologs data. As in Supplementary Table 1, the third column denotes the number of motifs found in the interaction network of 3,128 yeast proteins, obtained by counting subgraphs of two to five nodes. Our set of orthologs contains again all 678 proteins that have an orthologue in each proteome of human, mouse, fly, worm and *A. thaliana*. The natural conservation rate shows the fraction of the original yeast motifs that are evolutionary fully conserved. Accounting for the influence of false negative orthologues sequences, 10% (col. 5) and 20% (col. 4) of the orthologs, respectively, were randomly chosen, deleted and conservation rates of motifs thus calculated. To mimick the influence of false positive orthologs, the same procedure was repeated picking randomly 10% and 20 %, respectively, more proteins previously absent in the original set of orthologs (col. 6,7). Although the numbers of newly added or removed orthologs are significant, highly inter-connected motifs (triangles #3, squares #9 and pentagons #11) again remain the most overrepresented motifs.

#	Motifs	Number of *Yeast* Motifs	Natural Conservation Rate	Random Conservation Rate	Conservation Ratio
1		8,692	12.74 %	4.58 %	2.78
2		158.912	4,64 %	1.10 %	4.34
3		2,310	15.19 %	1.07 %	13.82
4		3,580,496	0.66 %	0.28 %	2.41
5		1,621,883	2.50 %	0.19 %	12.83
6		10,546	8.73 %	0.18 %	47.23
7		112,492	5.73 %	0.14 %	41.23
8		7,662	14.71 %	0.14 %	103.39
9		739	13.67 %	0.28 %	77.69
10		17,118	5.03 %	0.09 %	134.68
11		199	12.27 %	0.00 %	inf

Supplementary Table 3: The evolutionary conservation of non-complex motifs. The third column denotes the number of motifs obtained by counting all subgraphs of two to five nodes found in the yeast protein interaction network of 3,143 proteins and 8702 interactions. This network has been cleaned from interactions which originate from annotated complexes of the Swissprot database. Thus, only 31 proteins but 564 interactions were removed from the original network. As set of orthologs we choose 678 proteins that have an ortholog in each of the five studied higher eukaryotes, and identified all motifs for which each component belongs to this evolutionary conserved protein subset. The natural conservation rate shows the fraction of the original yeast motifs that are evolutionary fully conserved, i.e., each of their protein components belong to the 678 orthologs of the list. The random conservation rate denotes the fraction of motifs which are found to be fully conserved for a random ortholog distribution. The last column denotes the ratio between the natural and the random conservation ratios, indicating that all motifs are highly conserved.

References

1. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).

2. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sc.i U.S.A.* **98**, 4569-4574. (2001).

3. Von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein protein interactions. *Nature* **417**, 399-403 (2002).

4. Hazbun, T.R. & Fields, S. Networking proteins in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 4277-4278. (2001).

5. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.* **30**, 31-34 (2002).

6. Costanzo, M.C. *et al.* The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucl. Acids Res.* **28**, 73-76 (2000).

7. Schwikowski, B., Uetz, P. & Fields, S. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257-1261 (2000).

8. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**, 303-305 (2002).

9. Tatusov, R.L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucl. Acids Res.* **29**, 22-28 (2001).

10. Remm, M., Storm, C.E.V. & Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041-1052 (2001).

11. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.* **31**, 365-370 (2003).