Supplementary Online Material for: Redefining Edges: Relating 3D Structures to Protein Networks Provides Evolutionary Insights

Philip M. Kim, Long J. Lu, Yu Xia, and Mark Gerstein

Materials and Methods

Interaction data sources

Interaction data was gathered from a number of different published high-throughput datasets and databases (*S1-6*). Independent genomic features (Biological function, mRNA co-expression, cellular localization and essentiality) and Bayesian integration were used to eliminate noise from the dataset following the approach by Jansen et al. (*S7*). The resulting filtered interaction network is available as supplementary table S3.

Domain mapping and structural annotation of interactions

We used iPfam as a database of structurally characterized interactions, yielding a network of interacting Pfam domains (*S8*). We mapped all interacting Pfam (*S9*) domains onto yeast ORFs using the HMM-profile based mapping available in the Pfam mysql tables. Specifically, we used the "in_full" tag, meaning that the corresponding sequence appeared in the full profile alignment and the domain assignment is considered reliable by the Pfam team (*S9*). The vast majority (>95%) of these domain to sequence assignments had E-values<0.000001. These sequences are likely to have a similar structure as the solved structure that was mapped in iPfam (via E-MSD (*S10*) and Pfam). Then we checked for each interaction in the interaction network from above for interacting Pfam domains, keeping only the interactions between two ORFs which both contained Pfam domains that are seen to interact in iPfam. The resulting dataset comprises an intersection of iPfam and the filtered interaction dataset from above. For each interaction, aside from the protein pair, we have knowledge both the Pfam domains involved in the interaction as well as the detailed atomic structure of the interaction. As the annotation with Pfam domains represents a substantial filtering, our dataset can be viewed as a network of high-confidence interactions.

3D structural exclusion to distinguish interfaces (For all interactions derived from Pfam mapping)

To distinguish between simultaneously possible and mutually exclusive interactions we used a principle of 3D structural exclusion. Simply put, if protein A interacts with two Proteins B_1 and B_2, and both interactions share the same physical interaction interface (i.e., interact with overlapping residues as seen from the domain in protein A that mediates the interaction with B_1 and B_2), protein A can not possibly interact with both B_1 and B_2 at the same time. In that case, we call the interactions between A and B_1 and B_2 mutually exclusive and protein A to have one binding interface. If proteins B_1 and B_2 interact with A through distinct interaction interfaces (i.e., no overlapping residues) both

interactions can occur simultaneously and we view them as simultaneously possible. Protein A is then viewed as having two binding interfaces. We also call interactions as simultaneously possible if protein A contains more than one of the domains that mediate the interaction between A and the partners (B_1 and B_2). In this case, it is quite possible that both B_1 and B_2 interacts with A at the same time, through the several mediating domains.

The platinum standard dataset

To supplement the Pfam derived dataset with a dataset of even higher confidence, we examined all the yeast protein complexes with a known 3D structure for direct physical contacts between the protein chains. We used a cutoff of molecular contacts (4.5 Angstroms) between at least 2 residues to call two chains interacting. Since this set represents a real set of physical contacts of high confidence, we refer to it as the platinum standard. However, it may still contain artifacts, such as non-biological crystal contacts and obviously the number of false negatives is very large. Note that the large ribosome subunit appears as several disjoint complexes as there is interspersed RNA such that many protein chains do not touch each other. We merged it with the interaction dataset obtained from Pfam mapping to obtain the structural interaction network (SIN). If an interaction between two proteins was derived from the Pfam mapping and is also contained in the platinum standard, we discarded the Pfam mapping and kept the platinum standard interaction. In the platinum standard, obviously, interfaces can be counted directly and all interactions derived from here are simultaneously possible. The platinum standard contributes about 20% of the interactions, whereas the other 80% are derived from structural homologous mapped through iPfam.

Correlations with genomic features

GO biological process, molecular function and cellular component were taken from the SGD Lite annotation (*S11*). Co-expression correlation was calculated based on the rosetta compendium dataset (*S12*). Protein essentiality was measured based on the essential gene list from SGD (*S11*) and evolutionary rates (dN/dS ratios) were taken from the adjusted values given by Wall et al. (*S13*). The difference in protein essentiality and evolutionary rate between the average of our dataset (32% and 0.047, respectively) and the average protein in the proteome (~18% for protein essentiality (*S14*) and 0.077 for evolutionary rate) likely stem from the fact that in our dataset, every protein has at least one interaction partner. All p-values were calculated using the Wilcoxon ranksum test, testing the probability of two distributions having the same median.

Correlations with evolutionary rate controlling for expression abundance

We took the Codon adaptation index (CAI) as a good (and laboratory independent approximation) for mRNA abundance (*S13*). First we find that the average CAI is virtually identical for singlish-interface and multi-interface hubs (0.34 and 0.341 respectively, wilcoxon p-value of difference is 0.2). Then we chose a range of CAI's where the correlation between CAI and dN/dS ratio was small (chose $0.3<CAI<0.35$). In this range the Spearman correlation almost disappears (-0.1, p-value of 0.32). We recomputed previous calculations in this range of CAIs only.

Surface area calculations

Surface area calculations were carried out for the platinum standard dataset only, using the POPs program (*S15*). The interface surface area was calculated by successively deleting the two chains of the interaction from the PDB structure and calculating the total surface area, substracting it from the total. For the total accessible surface area used in interfaces, the interface surface area for all mutually exclusive interactions was added up. The adjusted total interface surface area was calculated by substracting the molecular weight from the total interface surface area, and was subsequently binned into 3 bins.

Measurement of paralogy

As a measure of paralogs, membership in the same eukaryotic cluster of orthologous groups (KOG (*S16,S17*)) was used for assignment of paralogs. For calculating the proportion of paralogs among protein pairs with the same partner and the same interface the entire SIN was used. For calculating the proportion of paralogs among protein pairs with the same partner and different interface, only the platinum standard was used to avoid potential problems that stem from the interface assignment process above, especially with multidomain proteins.

Supplementary Results

Diversity and Bias of the SIN

While the SIN is small compared to some of the large interaction databases such as DIP (*S6*), it is of similar size to early networks (*S1*) and other datasets in which we have high confidence in, such as the Filtered Yeast Interactome (FYI) (*S18*).

A total of 473 different domains are present in proteins in the SIN. The majority of those were used in structural assignments of interfaces (287). The size (of those which were used in interface assignments) ranges from very small domains (18 amino acid residues – a tandem hexapeptide repeat that form left-handed helices, or 24 amino acids – a zinc finger) to very large ones (795 residues – a myosin head). While this indicates that the SIN is representative of the yeast proteome, it is to assume from the selection based on interaction that have been resolved structurally that it is depleted in naturally disordered regions. A detailed listing of domains present is given in Table S1. Note, that in addition to these, the platinum standard contributes additional domains not listed here.

In total 212 different GO biological process categories are represented by proteins in the SIN. A detailed listing is given in Table S2.

Some networks are dominated by few large complexes, partly due to the "matrix model" (all proteins in one complex interact with each other). This is much less the case in the SIN, since we only focus on real physical interactions. In particular, e.g., the over one quarter of the FYI (26%) was due to interactions from the "well-known" complexes (i.e., RNA-polymerase, Proteasome and Ribosome), whereas in the SIN it is less than 18% – in the newer SIN (see below) it is less than 10%).

Since only interactions with structurally resolved interfaces are in the SIN, it is likely that disordered regions are underrepresented. Future work will combine structural methods we have used here with approaches that include disordered regions.

The entire SIN is given as supplementary Table S3.

New and updated version of the SIN

While this manuscript was in preparation and under review, a number of new interaction datasets appeared in the literature (*S19-22*). Combining this new data with our existing data does not change any of the conclusions in our work, in fact, most of our findings, e.g. the differences between singlish- and multi-interface hubs listed in Table 3 get even more significant. The updated SIN now contains 1178 nodes and 2195 edges, of which roughly half (1196) are classified as mutually exclusive. The new and significantly larger version of the SIN is available at:

http://networks.gersteinlab.org/structint

Examples of singlish and multi-interface hubs

Detailed listings of all singlish- and multi-interface hubs are given in Table S6. A good example of a multi-interface hub is Pre7p: It has 11 interaction partners and 7 interfaces. It is an integral member of the proteasome and therefore has many simultaneously possible interactions. Due to its role in an important and large complex it is essential for the cell. Another example is Arp2p has 5 interaction partners and 3 interfaces. It is a central part of the Arp2/3 complex, which has 7 subunits and is required for the motility of actin patches. It is hence an essential protein for the cell.

A good example for a singlish-interface hub is Snf1p. In the SIN, it has 2 interfaces and 5 interactions partners. It is quite likely that one of its interfaces is regulatory and one is catalytic. It is a protein kinase and probably interacts only transiently with most of it targets. Furthermore, despite of its role as a central regulator in a variety of cellular processes, e.g. carbon metabolism, sporulation or peroxisome biogenesis, it is an essential protein; the *SNF1* null-mutant is viable.

Supplementary Tables

Table S1: Pfam Domains present in the SIN. See attached file.
Table S2: GO biological process categories present in the SIN. See attached file.
Table S3: The structural interaction network (SIN). See attached file.

Table S4: Correspondence of simultaneously possible interactions and mutually exclusive interactions to Munich Information Center for Protein Sequences (MIPS) complexes. It was measured for which fraction of simultaneously possible and mutually exclusive interactions both interaction partners are members of the same MIPS complex (*S23*).

Interaction type	Both partners share same MIPS complex
Simultaneously possible	51%
Mutually exclusive	11%

Table S5: Difference of singlish- and multi-interface hubs with respect to genomic features when changing the degree-cutoff for hubs. Note that some of the p-values are not significant when changing the cutoff, however, the difference remains. This is due to the fact that the dataset is fairly small – when doing the same calculation on the larger new version of the SIN (SIN 2.0), all these differences remain significant.

Degree cutoff	Genomic feature	Whole dataset	p-value (All-singlish)	Singlish-interface hubs only	p-value (Singlish-multi)	Multi-interface hubs only
≥3	Protein essentiality	32.3%	0.4	33.9%	<0.01	52.3%
	Expression correlation	0.20	0.9	0.20	0.05	0.24
	Evolutionary rate	0.047	0.9	0.490	<0.01	0.300
≥4	Protein essentiality	32.3%	0.8	33.3%	<0.01	58.2%
	Expression correlation	0.20	0.5	0.17	0.01	0.23
	Evolutionary rate	0.047	0.7	0.420	<0.01	0.029
≥5	Protein essentiality	32.3%	0.9	31.8%	<0.01	64.9%
	Expression correlation	0.20	0.3	0.17	<0.05	0.25
	Evolutionary rate	0.047	0.5	0.051	<0.01	0.029
≥6	Protein essentiality	32.3%	1.0	35.0%	<0.01	68.0%
	Expression correlation	0.20	0.3	0.14	0.08	0.22
	Evolutionary rate	0.047	0.6	0.048	<0.05	0.028
≥7	Protein essentiality	32.3%	0.8	36.2%	<0.01	67.3%
	Expression correlation	0.20	0.3	0.16	0.10	0.22
	Evolutionary rate	0.047	0.7	0.048	0.08	0.029

Table S6: Detailed listing of all singlish- and multi-interface hubs. See attached file.

Table S7: Correspondence of multi-interface hubs and singlish-interface hubs to date and party hubs. As can be seen, multi-interface hubs correspond mostly to party-hubs, whereas singlish-interface hubs correspond mostly to date-hubs. One reason that the overall overlaps are fairly small is that the overlap between the SIN and the FYI is small to begin with.

Hub type	Party Hubs	Date Hubs
Multi-interface hubs	24%	4%
Singlish-interface hubs	5%	16%

Table S8: Correspondence of multi-interface hubs and singlish-interface hubs to MIPS complexes. As expected, multi-interface hubs correspond strongly to central members of MIPS complexes and singlish-interface hubs do so only rarely.

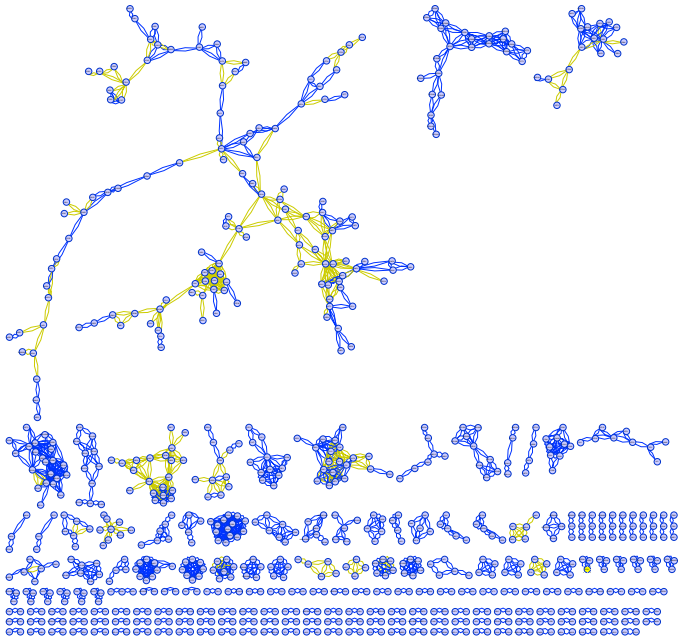
Hub type	Part of a MIPS complex
Multi-interface hubs	84%
Singlish-interface hubs	13%

Table S9: Evolutionary rates at different positions in the protein structure. Buried core refers to residues in the core of the protein that is not solvent accessible. Surface - interface refers to residues on the protein surface which are involved in interactions with other proteins. Surface - exposed refers to residues on the protein surface which are not involved in interactions and are therefore exposed. Measured were the number of mutations per site based on protein sequence alignments of *S. Cerevisiae* with *S. Bayanus*.

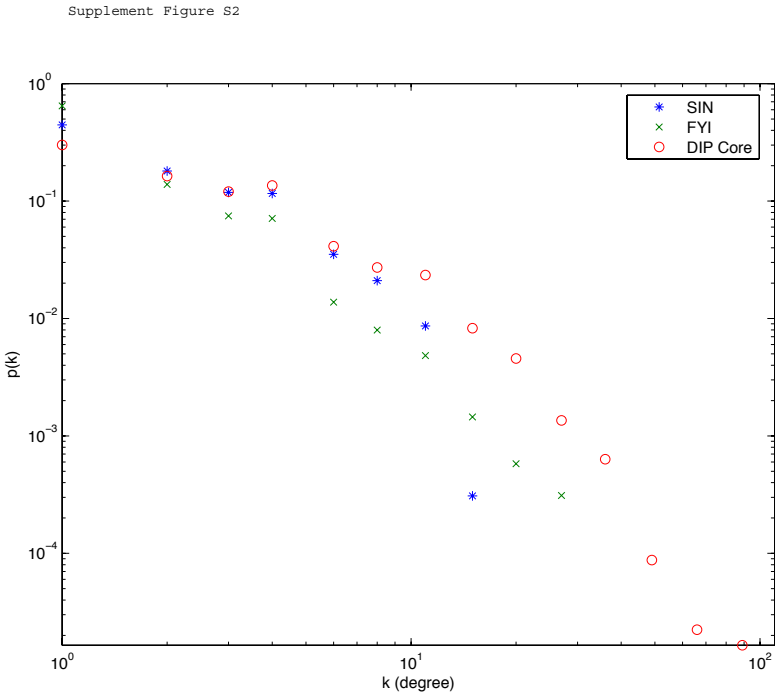
Site	Mutations/site (Scer-Sbay)
Buried core	0.058
Surface - interface	0.065
Surface - exposed	0.124

Supplementary figures:

Supplement Figure S1

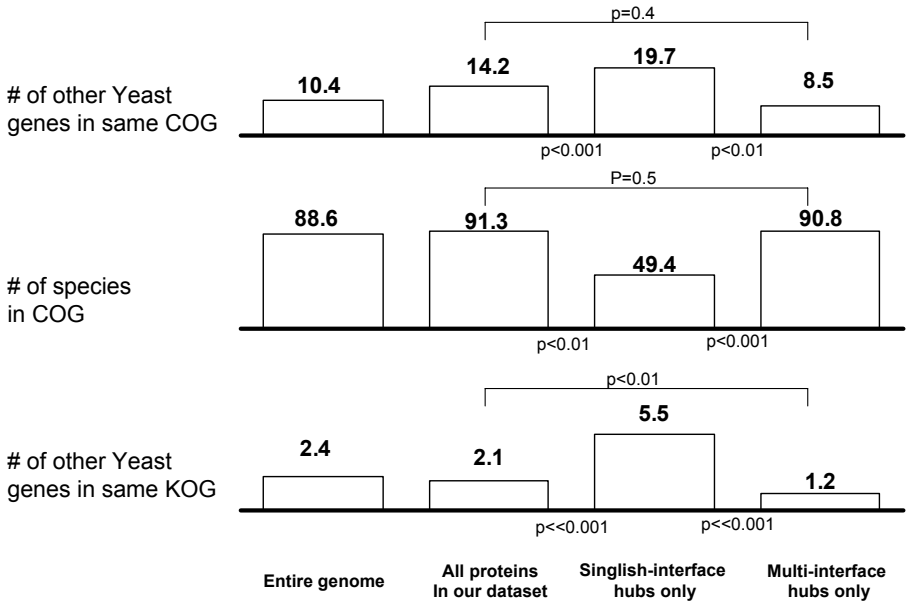


Supplement figure S1: The structural interaction network (SIN) is shown. Each yeast ORF is depicted as a node and each interaction as an edge. Each edge has been annotated with a corresponding PDB structure (the platinum standard) or two Pfam domains, with a corresponding PDB structure of the interaction. Simultaneously possible interactions are shown in blue and mutually exclusive interactions in yellow. The resulting network consists of 873 nodes and 1269 interactions. 530 of the interactions are from direct observations in the PDB, and the rest of the interactions are from Pfam mappings.



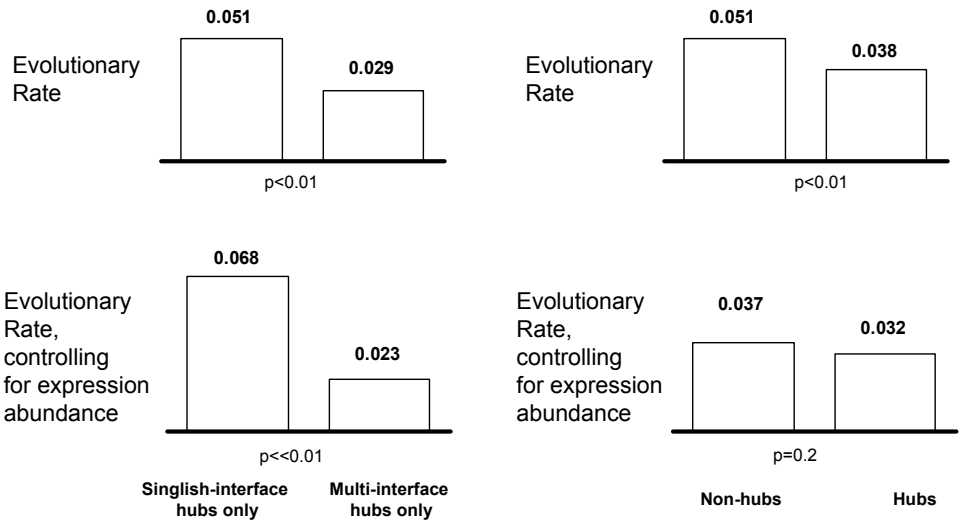
Supplement figure S2: The degree distribution is shown for our structural network (SIN), the Filtered Yeast Interactom (FYI (*S18*)) and the DIP core (*S24*), two current high-confidence interaction datasets. For the degree calculation each unique interaction is counted once. As can be seen, the degree distribution has a markedly shorter tail for the SIN. Many of the interactions observed previously (those that are not false positives) are likely due to a "matrix-model" of protein complexes, in which all proteins belonging to the same complex are annotated as interacting, even if they don't physically touch each other. In particular the maximum degree in the SIN is 14, reflecting physical constraints on direct physical interactions. Although it is theoretically possible to have many more mutually exclusive interactions binding at the same interface, their actual number appears to be limited.

Supplement Figure S3



Supplement figure S3: Evolutionary measures for singlish-interface and multi-interface hubs. Singlish-interface hubs have more yeast genes that are in the same COG or KOG, suggesting that they were more often duplicated. By contrast, multi-interface hubs are less likely to have been duplicated. Also, singlish-interface hub proteins appear to have been less conserved, as measured by the number of species in a given COG.

Supplement Figure S4



Supplement figure S4: Evolutionary rates (dN/dS rates) for singlish-interface, multi-interface hubs and non-hubs and hubs. As can be seen, for both the degree and the number of interfaces (controlling for the degree), the evolutionary rate drops significantly at higher levels. However, when controlling for the expression level (Limiting the CAI to the interval [0.3,0.35], where correlation between CAI and dN/dS rates almost disappears), the dependence on the degree disappears (as has been pointed out before (*S25*)), but the dependence on the number of interfaces remains, in fact, the difference between singlish-interface hubs and multi-interfaces hubs becomes stronger.

Supplement references:

S1. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
S2. H. W. Mewes *et al.*, *Nucleic Acids Res* **30**, 31 (2002).
S3. T. Ito *et al.*, *Proc Natl Acad Sci U S A* **97**, 1143 (2000).
S4. Y. Ho *et al.*, *Nature* **415**, 180 (2002).
S5. A. C. Gavin *et al.*, *Nature* **415**, 141 (2002).
S6. I. Xenarios *et al.*, *Nucleic Acids Res* **30**, 303 (2002).
S7. R. Jansen *et al.*, *Science* **302**, 449 (Oct 17, 2003).
S8. R. D. Finn, M. Marshall, A. Bateman, *Bioinformatics* **21**, 410 (Feb 1, 2005).
S9. A. Bateman *et al.*, *Nucleic Acids Res* **30**, 276 (Jan 1, 2002).
S10. A. Golovin *et al.*, *Nucleic Acids Res* **32**, D211 (Jan 1, 2004).
S11. L. Issel-Tarver *et al.*, *Methods Enzymol* **350**, 329 (2002).
S12. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).
S13. D. P. Wall *et al.*, *Proc Natl Acad Sci U S A* **102**, 5483 (Apr 12, 2005).
S14. G. Giaever *et al.*, *Nature* **418**, 387 (Jul 25, 2002).
S15. L. Cavallo, J. Kleinjung, F. Fraternali, *Nucleic Acids Res* **31**, 3364 (Jul 1, 2003).
S16. C. von Mering *et al.*, *Nucleic Acids Res* **31**, 258 (Jan 1, 2003).
S17. R. L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (Sep 11, 2003).
S18. J. D. Han *et al.*, *Nature* **430**, 88 (Jul 1, 2004).
S19. A. C. Gavin *et al.*, *Nature* (Jan 22, 2006).
S20. N. J. Krogan *et al.*, *Nature* **440**, 637 (Mar 30, 2006).
S21. T. Reguly *et al.*, *J Biol* **5**, 11 (2006).
S22. C. Stark *et al.*, *Nucleic Acids Res* **34**, D535 (Jan 1, 2006).
S23. H. W. Mewes *et al.*, *Nucleic Acids Res* **34**, D169 (Jan 1, 2006).
S24. C. M. Deane, L. Salwinski, I. Xenarios, D. Eisenberg, *Mol Cell Proteomics* **1**, 349 (May, 2002).
S25. J. D. Bloom, C. Adami, *BMC Evol Biol* **3**, 21 (Oct 2, 2003).