Molecular Systems Biology (2006) doi:10.1038/msb4100103
© 2006 EMBO and Nature Publishing Group All rights reserved 1744-4292/06
www.molecularsystemsbiology.com
Article number: 66

Biological context networks: a mosaic view of the interactome

John Rachlin[1,2,*], Dikla Dotan Cohen[2], Charles Cantor[3,4,5] and Simon Kasif[2,3,6]

[1] Department of Computer Science, Boston University, Boston, MA, USA, [2] Center for Advanced Genomic Technologies, Boston University, Boston, MA, USA, [3] Department of Biomedical Engineering, Boston University, Boston, MA, USA, [4] Center for Advanced Biotechnology, Boston University, Boston, MA, USA, [5] SEQUENOM Inc., San Diego, CA, USA and [6] Children's Hospital Boston, Boston, MA, USA
[*] Corresponding author. Department of Computer Science, Boston University, 111 Cummington Ave, Boston, MA 02215, USA. Tel.: +1 617 921 9669; Fax: +1 617 353 4814; E-mail: rachlin@bu.edu

Received 30.1.06; accepted 22.9.06

Network models are a fundamental tool for the visualization and analysis of molecular interactions occurring in biological systems. While broadly illuminating the molecular machinery of the cell, graphical representations of protein interaction networks mask complex patterns of interaction that depend on temporal, spatial, or condition-specific contexts. In this paper, we introduce a novel graph construct called a biological context network that explicitly captures these changing patterns of interaction from one biological context to another. We consider known gene ontology biological process and cellular component annotations as a proxy for context, and show that aggregating small process-specific protein interaction sub-networks leads to the emergence of observed scale-free properties. The biological context model also provides the basis for characterizing proteins in terms of several context-specific measures, including 'interactive promiscuity,' which identifies proteins whose interacting partners vary from one context to another. We show that such context-sensitive measures are significantly better predictors of knockout lethality than node degree, reaching better than 70% accuracy among the top scoring proteins.

Molecular Systems Biology 28 November 2006; doi:10.1038/msb4100103
Subject Categories: metabolic and regulatory networks
Keywords: bioinformatics; biological context; network models; PPI networks; scale-free networks

Introduction

Graphs and their variants are the foundation for modeling complex biological systems. Graph topology reveals the basic properties of connectivity, robustness, modularity, hierarchical structure, and other properties, enabling identification of protein complexes or functional modules (Segal *et al*, 2003; Spirin and Mirny, 2003), and serves to aid whole-genome annotation efforts (Hartwell *et al*, 1999; Marcotte *et al*, 1999; Zheng *et al*, 2002; Letovsky and Kasif, 2003; Karaoz *et al*, 2004). Cross-species comparisons of conserved sub-networks reveal broad classes of conserved networks (Milo *et al*, 2004; Sharan *et al*, 2005) and recurring motifs (Vazquez *et al*, 2004). Biological networks are also of commercial interest as an aid to drug target discovery or for predicting toxic side effects, and they are at the heart of pharmaceutical initiatives focused on integrating and mining pathways data sets (Gardner *et al*, 2003; Hood *et al*, 2004; di Bernardo *et al*, 2005).

Biological interaction networks are often obtained by high-throughput detection assays (Giot *et al*, 2003; Rual *et al*, 2005) or inferred from literature surveys (Mishra *et al*, 2006). As a result, they represent a high-level integrated summary of a large number of interactions inferred from many biological contexts. However, representing the interactome as a static biological network is akin to a long-exposure photograph that can mask more complex patterns of activation across multiple processes, cellular locations, and time. Conclusions drawn from the full network's topology may be compromised by these inherent limitations. A central goal of systems biology research is to elucidate the underlying patterns of interaction, in an effort to obtain more realistic and predictive models of the cell (Ideker *et al*, 2001; Hood, 2003). This has prompted the development of a broad range of graphical representations coupled with mathematical equations intended to model cellular dynamics. By contrast, protein–protein interaction (PPI) networks are typically represented as a standard undirected graph where vertices correspond to individual proteins and edges connect pairs of interacting proteins. In this paper, we propose an intermediate-level model, called a 'biological context network,' in which we label proteins with contextual information about the protein and activate protein interactions as specified by the succinct biological program associated with the network. In its simplest form, the program activates an edge, whenever two interacting proteins are in a shared contextual state, and otherwise assumes that the interaction has been inactivated. The biological context network model enables one to view the interactome as a mosaic of overlapping sub-networks, each associated with specific contexts or conditions, and to further characterize changes in topology from one context to another. For example,

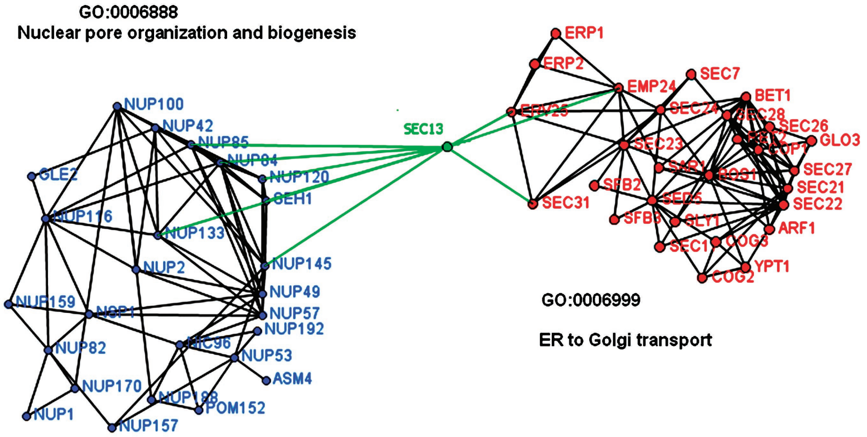


Figure 1 The local context networks for Sec13 with respect to two of its current GO biological process annotations—GO:0006888, nuclear pore organization and biogenesis, and GO:0006999, ER to Golgi transport—highlighting Sec13's association with both the nuclear pore complex and the ER. Sec13 is an example of a protein whose interacting partners vary from one process context to another. We characterize such proteins as 'interactively promiscuous.' The shuttling of Sec13 between the nucleus and the cytoplasm is believed to play a cross-functional regulatory role (Enninga *et al*, 2003).

in Figure 1, we show the context-specific sub-networks in the local neighborhood of the protein Sec13, highlighting its association with both the nuclear pore complex and the endoplasmic reticulum (ER) (Enninga *et al*, 2003).

Previous research investigating condition-specific sub-networks in regulatory networks has proven useful in identifying static and transient hubs (Luscombe *et al*, 2004). In this paper, we use gene ontology (GO) (Harris *et al*, 2004) biological process and cellular component annotations as representatives of putative contextual assignments, and analyze the PPI network of the yeast *Saccharomyces cerevisiae*. We investigate the topological changes that occur from one context to another and focus specifically on identifying proteins whose interacting partners are highly context-dependent.

We generate a context-specific sub-network as follows: from the original network, we extract any proteins associated with a particular context (i.e., having a particular GO term) and then include any interconnecting edges from the full PPI network that occur between any pair of nodes in this resulting protein subset. It is important to emphasize that the resulting edges do not signify that each activated interaction definitively occurs within the specified context, but merely that the interactions are *consistent* with the contextual labels of the interacting proteins. Nevertheless, we believe that the explicit modeling of contextual information in biological networks, even in an approximate manner, may offer a new perspective on the observed scale-free topologies in PPI networks and provide novel characterization of individual proteins within a changing network topology.

It has been widely observed that a broad range of social, technological, and biological networks are scale-free, characterized by a power-law degree distribution where a few 'hub' proteins have many interacting partners, whereas most proteins have very few (Barabasi and Albert, 1999; Amaral *et al*, 2000; Barabasi and Oltvai, 2004; Han *et al*, 2004; Yook *et al*, 2004). For PPI networks, high-degree 'hubs' are more likely to be essential for the viability of the organism. In this paper, we provide evidence that a power-law distribution, while clearly evident in the aggregate experimental PPI data, is plausibly an artifact of the aggregation of interactions across multiple process-specific contexts. This observation suggests that paths connecting disparate protein pairs may be substantially impacted by intervening contextual differences. We show, for example, that aggregating about 100 small leaf-term sub-networks reconstitutes a scale-free network (R^2=0.88). In order to better gauge the rapidity with which scale-free topologies emerge through aggregation, we have also simulated this effect on the aggregation of random Erdös-Rényi networks, each representing a particular shared context, but subject to the constraint that the number of contextual labels (or annotations) per protein follows a power-law distribution—a fact we confirm for a wide range of species.

The analysis of context-specific sub-networks derived using the biological context network model provides, in addition, a basis for characterizing proteins with respect to several context-specific measures. These include the following:

- *Context degree*: The degree of a node, considering only those interacting partners that share at least one context (annotation), while taking into account transitive closure within the GO annotation hierarchy. An edge, thus, is preserved if the neighbor has at least one annotation in common with the protein in question, or if it contains at least one annotation more specific than an annotation of the protein. Equivalently, context degree is the number of unique edges occurring in at least one context-specific sub-network.
- *Context mutual information*: Measures the degree to which the annotations of neighboring proteins are correlated.
- *Interactive promiscuity*: Measures the variability of annotations (contexts) among a protein's interacting partners in an effort to identify those proteins likely to play a cross-contextual 'linking' role. We emphasize that an interactively promiscuous protein is not necessarily promiscuous in the sense of being involved in multiple processes, or being merely highly annotated. Interactive promiscuity is specifically concerned with changes in the network topology from one context to another. By contrast, a protein whose

interacting partners are stable from one context to another is 'interactively conserved.'

We provide formal definitions for context mutual information and interactive promiscuity below. Interestingly, we find that the top-ranked proteins with respect to each of these context-specific measures are highly enriched in essential proteins and these measures provide a significantly improved predictor for knockout lethality than the static measure of the degree computed from the original 'context-free' network.

The distinction between interactively promiscuous and interactively conserved proteins is similar in conception although different in practice from the classification of hubs as either date hubs (interacting at different times and/or places) or party hubs (interacting at the same time and place) introduced by Han *et al* (2004). The distinction is addressed in greater detail below. In part, this difference stems from the fact that the date/party hub distinction is based on the correlation of expression patterns (or lack thereof), whereas our focus is on the existing GO biological process and cellular component annotations as putative process-specific or location-specific contexts.

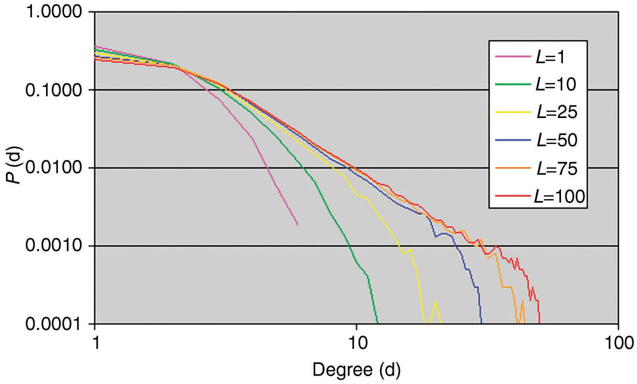


Figure 2 A simulation of the aggregation of random (Erdös-Rényi) graphs showing rapid reconstitution of a scale-free degree distribution. The simulation involved a set of random labels (contexts), *L*, distributed across *N*=1000 nodes, subject to the additional condition that the number of labels per node follows a power-law distribution. Displayed results are the average of 100 trials, showing degree distribution after aggregating *L*=1, 25, 50, 75, and 100 labels. In our simulation, we assume that two nodes sharing a given context have a fixed probability of interaction (*P*=0.05), thus any context-specific sub-network is a random (Erdös-Rényi) graph.

Results

Aggregation of context-specific sub-networks rapidly reconstitutes a scale-free network

A graphical representation of the yeast PPI network represents proteins as nodes and physical interactions between two proteins as edges. We examined a PPI data set from the database of interacting proteins (DIP) containing 15 429 interactions among 4741 proteins. It exhibits a scale-free topology (γ=1.79, R^2=0.88). It is widely appreciated that the static PPI network view of an organism's interactome simplifies what is otherwise a complex dynamic system governed by context-specific pattern interactions mediated by the activation of specific cellular processes and regulated by the transcriptional machinery of the cell. We show by simulation that the aggregation of random (Erdös-Rényi) networks can reconstitute a scale-free topology when the number of contexts per protein is itself characterized by a power-law distribution. We also demonstrate that the smallest (context-specific) sub-networks generated from the DIP PPI network will also form a scale-free network in aggregate. The small size of these networks (<14 edges per network) precludes a reliable characterization of their degree distribution, but demonstrates, nevertheless, that a seemingly fragmentary collection of context-specific sub-networks is sufficient to reconstitute a power-law degree distribution when context is ignored.

In our simulation (Figure 2), we constructed a random biological context network with *N*=1000 nodes. Each node was associated with one or more contextual states. Particular context labels were assigned randomly from a pool of labels, **L** (where |**L**|=100). We assumed that two proteins sharing a particular context interact within that context with fixed probability, *P*. We further assumed that the number of proteins having *c* contexts, N_c, is subject to a power-law distribution ($N_c \sim c^{-\gamma}$). This latter assumption roughly holds for a broad range of species, using the GO biological process annotations as representative of a specific contextual role (Supplementary Figure S1). The resulting context-specific sub-networks are random Erdös-Rényi graphs, as constrained by the fixed interaction probability employed in our simulation. Our aggregation results were averaged over 100 random trials. Aggregation of these networks rapidly reconstitutes a scale-free distribution. See Materials and methods section for detailed simulation procedures.

We recognize that this simulation has certain inherent limitations. For example, it is clear that the highest degree nodes in the aggregate network will generally correspond to those proteins that are most highly annotated, as these are the nodes that occur in the most context-specific sub-networks. In the yeast PPI network, hubs are often associated with protein complexes having a limited number of annotations. Indeed, we find no correlation in the DIP network (R^2=0.07) between the number of interacting partners and the number of annotations that the protein contains (Supplementary Figure S2). This observation is consistent with previous studies demonstrating a lack of correlation between the amount of information known about a gene and its centrality in a biological network (Hoffmann and Valencia, 2003). Furthermore, our simulation does not attempt to model the formation of complexes or other modular features of real biological networks. Nevertheless, the simulation demonstrates the plausibility of reconstituting a scale-free topology as an aggregate of non-scale-free context-specific sub-networks. We have found this effect to be independent of network size, number of context labels, the power-law scaling coefficient, or interaction probability (Supplementary Figure S3).

We next considered the extent to which aggregating process-specific sub-networks within the DIP network would reconstitute the inherent scale-free topology of the full PPI network. As explained earlier, a process-specific sub-network retains proteins having the corresponding GO biological process annotation, or a more specific term, and any interactions occurring between them from the full network. The resulting sub-network is an approximation of the cellular interactions associated with a particular process. Each edge is *consistent* with the GO biological process of interest (because all nodes are associated with the process), but it is nevertheless possible that specific interactions are activated in some other shared context or are merely the result of experimental error.

As GO annotations are hierarchical, high-level terms may retain large numbers of nodes and a substantial fraction of the original network (Supplementary Figure S4). For example, the high-level term *Protein Biosynthesis* (GO:0006412) produces a sub-network containing 253 proteins and 236 edges. Excluding 54 zero-degree proteins, the network is already inherently scale-free (R^2=0.86). It is not surprising that sub-networks corresponding to high-level terms exhibit a scale-free topology, because they extract a substantial fraction of the original network.

We identified 384 GO biological leaf terms occurring in the DIP PPI network for yeast, yielding a non-empty sub-network projection (>0 nodes). Of these, 137 contain at least one edge. Figure 3 plots these 137 leaf-term projections by the number of nodes and edges in the resulting sub-network, and also provides a sampling of the resulting networks. The small size of these sub-networks (average number of nodes=11.4, average number of edges=13.7) makes it problematic to characterize their degree distribution as scale-free or non-scale-free. We note, however, that the largest of these leaf-term sub-networks (GO:0006888—ER to Golgi transport) had 60 nodes and 123 interactions, and is moderately correlated with a scale-free distribution (R^2=0.75). As we wished to consider the effect of aggregating non-scale-free sub-networks, we eliminated 34 of the largest leaf-term sub-networks from further consideration. The remaining 103 sub-networks have no more than 21 proteins (mean=6.9 ± 4.4) and 13 interactions (mean=3.7 ± 3.5). When we simulate the cross-contextual aggregation of evidence by taking the union of these fragmentary leaf-term sub-networks, we find that the resulting

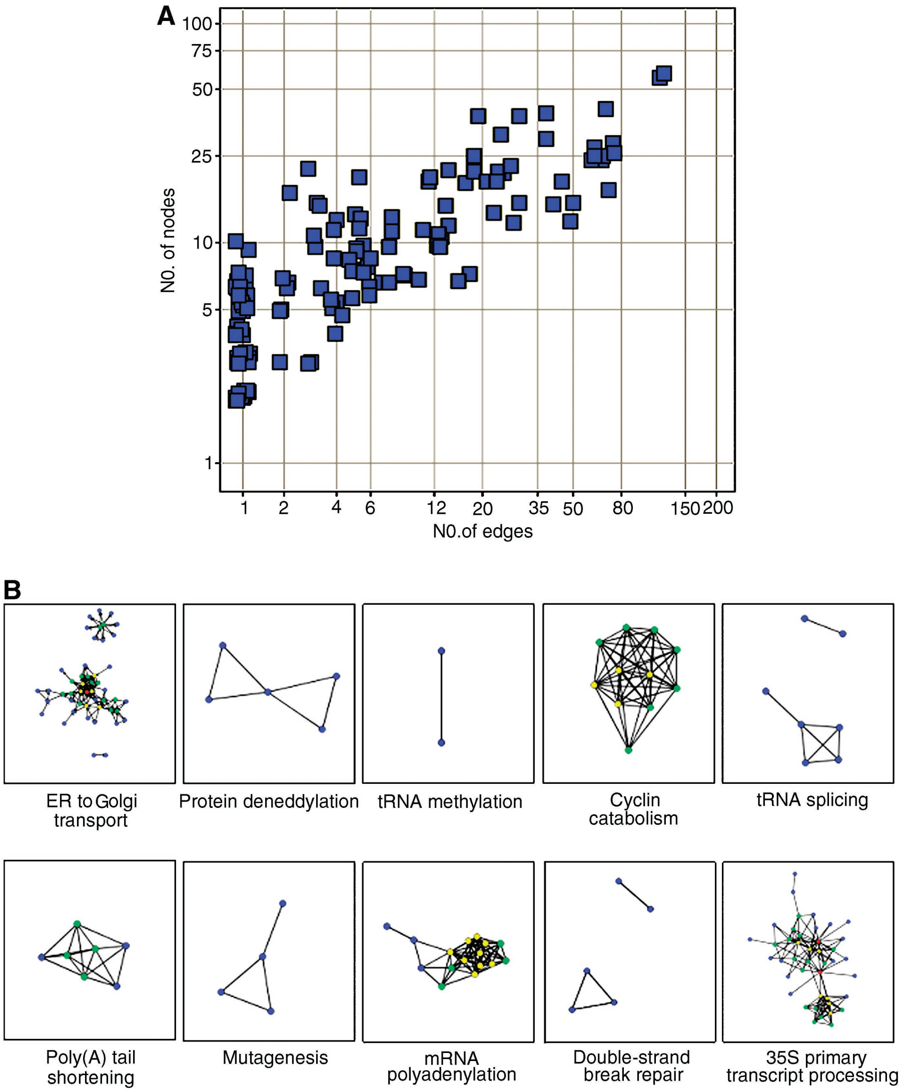
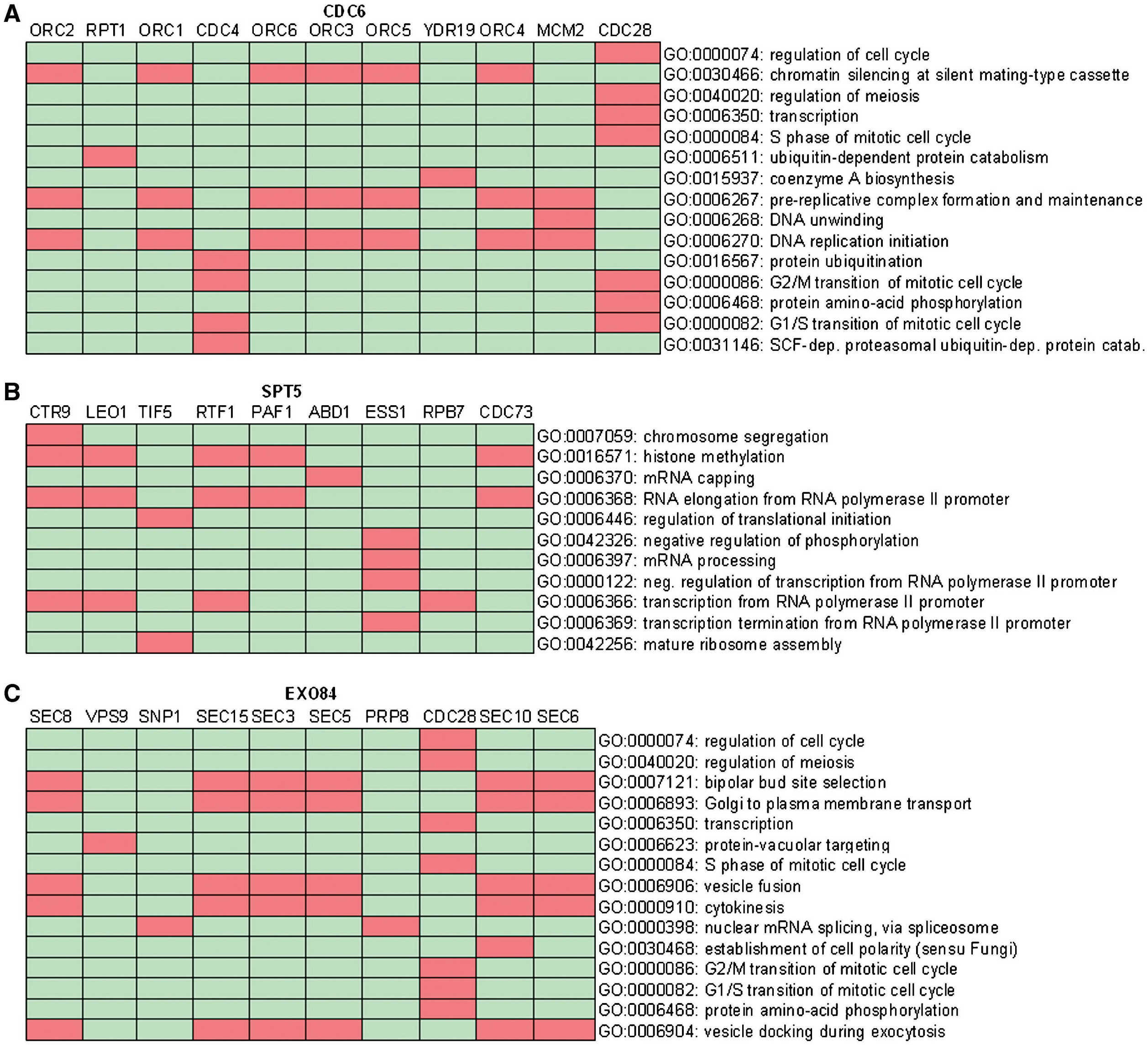


Figure 3 (**A**) Location of GO biological process leaf-term sub-networks in EDGE-NODE space, showing the variability in the size of the resulting projections. (**B**) The resulting sub-networks reveal a broad range of irregular distributions. Singleton nodes are excluded for clarity. Node color coding is by degree: 1–4 neighbors (blue), 5–9 neighbors (green), 10–14 neighbors (yellow), 15 + neighbors (red).



A CDC6

	ORC2	RPT1	ORC1	CDC4	ORC6	ORC3	ORC5	YDR19	ORC4	MCM2	CDC28
GO:0000074: regulation of cell cycle	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥
GO:0030466: chromatin silencing at silent mating-type cassette	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟩	🟥	🟩	🟩
GO:0040020: regulation of meiosis	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥
GO:0006350: transcription	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥
GO:0000084: S phase of mitotic cell cycle	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥
GO:0006511: ubiquitin-dependent protein catabolism	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
GO:0015937: coenzyme A biosynthesis	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩
GO:0006267: pre-replicative complex formation and maintenance	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟩	🟥	🟥	🟩
GO:0006268: DNA unwinding	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩
GO:0006270: DNA replication initiation	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟩	🟥	🟥	🟩
GO:0016567: protein ubiquitination	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩
GO:0000086: G2/M transition of mitotic cell cycle	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟥
GO:0006468: protein amino-acid phosphorylation	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥
GO:0000082: G1/S transition of mitotic cell cycle	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟥
GO:0031146: SCF-dep. proteasomal ubiquitin-dep. protein catab.	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩

B SPT5

	CTR9	LEO1	TIF5	RTF1	PAF1	ABD1	ESS1	RPB7	CDC73
GO:0007059: chromosome segregation	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
GO:0016571: histone methylation	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟩	🟥
GO:0006370: mRNA capping	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩
GO:0006368: RNA elongation from RNA polymerase II promoter	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟩	🟥
GO:0006446: regulation of translational initiation	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩
GO:0042326: negative regulation of phosphorylation	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0006397: mRNA processing	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0000122: neg. regulation of transcription from RNA polymerase II promoter	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0006366: transcription from RNA polymerase II promoter	🟥	🟥	🟩	🟥	🟩	🟩	🟩	🟥	🟩
GO:0006369: transcription termination from RNA polymerase II promoter	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0042256: mature ribosome assembly	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩

C EXO84

	SEC8	VPS9	SNP1	SEC15	SEC3	SEC5	PRP8	CDC28	SEC10	SEC6
GO:0000074: regulation of cell cycle	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0040020: regulation of meiosis	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0007121: bipolar bud site selection	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟥	🟥
GO:0006893: Golgi to plasma membrane transport	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟥	🟥
GO:0006350: transcription	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0006623: protein-vacuolar targeting	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
GO:0000084: S phase of mitotic cell cycle	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0006906: vesicle fusion	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟥	🟥
GO:0000910: cytokinesis	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟥	🟥
GO:0000398: nuclear mRNA splicing, via spliceosome	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟩	🟩
GO:0030468: establishment of cell polarity (sensu Fungi)	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩
GO:0000086: G2/M transition of mitotic cell cycle	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0000082: G1/S transition of mitotic cell cycle	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0006468: protein amino-acid phosphorylation	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩
GO:0006904: vesicle docking during exocytosis	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟥	🟥

Figure 4 Neighborhood-annotation matrices for three interactively promiscuous examples: (**A**) Cdc6, (**B**) Spt5, and (**C**) Exo84. Column headers include all neighbors having at least one GO biological process annotation. Rows correspond to particular GO annotations associated with these neighboring proteins. Red boxes indicate that the neighbor protein has the annotation explicitly or a more specific annotation in the GO biological process ontology.

aggregate network contains 578 proteins (333 having at least one edge) and 308 edges, and is well described by and highly correlated to a scale-free network (γ=−2.2, R^2=0.88), although it is better correlated with an exponential degree distribution (R^2=0.98), a distribution that is not at all characteristic of the original DIP network (R^2=0.41). The resulting network contains 369 unique annotations and 220 unique leaf-term annotations. The number of annotations per protein, as well as the number of *leaf-term* annotations per protein, exhibits a power-law distribution (R^2=0.90 and 0.91, respectively), consistent with our earlier simulations. This analysis demonstrates the overall tendency of small and highly specific sub-networks to reconstitute scale-free topologies in aggregate, although the precise process by which aggregate distributions ultimately achieve scale-freeness may involve transitional characteristics.

Using context to characterize the local network neighborhood

We define several measures used to characterize the local context network neighborhood. Intuitively, the context neighborhood is the set of contextual labels or annotations (e.g., GO biological process annotations) associated with the interacting partners of a particular protein. More formally, the context neighborhood of a protein is a matrix $M_{G\times K}$, where K is the number of interacting neighbors and G is the number of annotations occurring among the K interactors. For a given protein, M_{gk}=1, iff neighbor k has annotation g or a more specific annotation (to account for transitive closure in the GO term hierarchy), otherwise M_{gk}=0. The columns of the matrix thus correspond to the set annotated of interacting partners, whereas the rows define the set of annotations occurring at

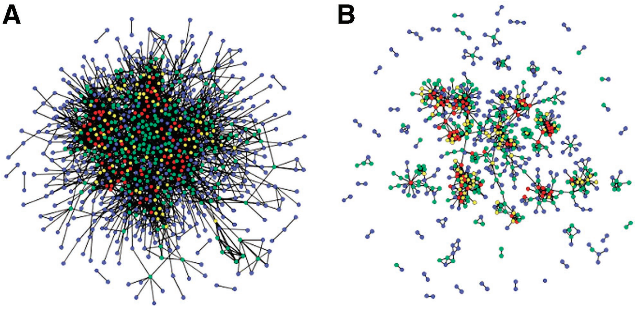


Figure 5 (**A**) DIP yeast sub-network of all 991 essential proteins. (**B**) Essential proteins having context degree >1. Node coloring is according to the degree of the protein in the full DIP network: 1–4 neighbors (blue), 5–9 neighbors (green), 10–14 neighbors (yellow), 15+ neighbors (red). Many of the essential proteins aggregate into clusters of essential protein complexes that are typically related to cell-cycle regulation and mRNA processing. As a result of the network's improved specificity, context degree is a better predictor for knockout lethality, although applicable only to annotated nodes.

least once among these interactors. Figure 4 presents the local context neighborhood matrices for three proteins: Cdc6, Spt5, and Exo84. Separate matrices were computed for each protein in the DIP network and separately generated for both GO biological process and GO cellular component terms. The resulting matrices are the basis of several context-specific measures described below.

Interactive promiscuity (IP) is the average pairwise Hamming distance across all annotations (row) vectors of the matrix:

$$\mathrm{IP} = \frac{2}{G(G-1)} \sum_{i,j}^{G} h(\vec{A}_i, \vec{A}_j)$$

where $\vec{A}_i$ denotes the bit vector formed from the ith annotation (row) of the matrix M and h is the Hamming distance function on two bit vectors, which measures the total number of bit changes ($0 \rightarrow 1$ or $1 \rightarrow 0$) across all vector positions. In this case, given two annotations, the Hamming distance function counts the number of neighbors having one annotation, but not the other. Interactively promiscuous proteins have different interacting partners in different contexts. Those proteins having the same interacting partners across multiple contexts are interactively conserved. We emphasize that interactive promiscuity is not equivalent to having multiple distinct annotations. Certain proteins may be richly annotated simply because they have been more thoroughly investigated. A multi-functional protein that maintains the same interactors across different GO terms is not considered to be interactively promiscuous by our definition. Thus, it is possible for a protein to be interactively promiscuous but have few annotations, or to be interactively conserved but have many annotations. For example, the proteins Nup170, Gle2, Tmp1/Tmp2, Rga2, and Elm1 have the highest number of GO biological process annotations, but are not found to be interactively promiscuous. By contrast, the proteins SPP381, Atp1/Atp2, and Smd{1,2,3} are found to be highly interactively promiscuous, but have relatively few annotations. This distinction also applies to interactive promiscuity based on GO cellular compartment annotations, where we find, for example, that there are proteins found in multiple cellular locations that are not, however, considered interactively promiscuous using GO cellular compartment as the basis for computing interactive promiscuity. In general, we find that interactive promiscuity based on GO biological process and GO cellular component annotations are well correlated (R^2=0.72; see Supplementary Figure S5a). Complete protein-by-protein statistics are provided in the supplement.

Context degree is the number of neighbors of a given protein having at least one shared context. If GO annotations are the basis for defining context, then the interacting partner must have at least one annotation that is either identical to or a descendant of an annotation of the node whose context degree we are measuring. A node that is un-annotated, or whose neighbors are all un-annotated, or whose neighbors have no common annotations has a context degree of zero. The DIP sub-network of essential proteins for both all edges and context-verified edges only is provided for comparison in Figure 5. Not unsurprisingly, many context-verified edges are often associated with protein complexes.

Context mutual information (CMI), like interactive promiscuity, is based on the context neighborhood matrix, M. It is defined as the sum total pairwise mutual information across all neighbors (column vectors) of the matrix, and measures the degree to which annotation among pairs of interacting neighbors is correlated. Formally,

$$\mathrm{CMI} = \sum_{i,j}^{K} I(\vec{K}_i; \vec{K}_j)$$

where $\vec{K}_i$ is the ith column vector of the neighborhood matrix M, and I is mutual information defined as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

where $p(x, y)$ is the joint probability distribution, $p(X{=}x$ and $Y{=}y)$, and $p(x)$ and $p(y)$ are the marginal probability distribution functions, $p(X{=}x)$ and $p(Y{=}y)$ respectively, with $X{=}\{0,1\}$ and $Y{=}\{0,1\}$. Here, mutual information is computed by assuming values for x and y along each row position of the two column vectors $\vec{K}_i$ and $\vec{K}_j$. Unlike interactive promiscuity, context mutual information values based on GO biological process and GO cellular component contexts are not well correlated (R^2=0.32; Supplementary Figure S5b), although we have not yet determined definitively why this might be the case.

We also find in general that the number of interacting partners (degree) does not correlate well with interactive promiscuity (R^2=0.26), context mutual information (R^2=0.15), or even context degree (R^2=0.39) when these measures are computed using GO biological process annotations. When based on GO cellular component annotations, the correlation of context mutual information to degree is substantially higher (R^2=0.62), but not for other measures.

Examples

We conducted an informal literature survey pertaining to those proteins found to be highly interactively promiscuous. This survey was by no means comprehensive. We identified 10

Table I Examples of 'interactively promiscuous' proteins

Name	SGD	Degree	Context degree	Context mutual information	Interactive promiscuity	GO biological process annotations
SEC13	YLR208W	17	9	13.4	4.1	Nuclear pore organization and biogenesis, ER-associated protein catabolism, ER to Golgi transport, vesicle budding
ACT1	YFL039C	40	26	8.8	4.0	Mitochondrion inheritance, protein secretion, actin filament reorganization during cell cycle, exocytosis, sporulation (sensu Fungi), chronological cell aging, endocytosis, establishment of mitotic spindle orientation, response to osmotic stress, establishment of cell polarity (sensu Fungi), regulation of transcription from RNA polymerase II promoter, vacuole inheritance, vesicle transport along actin filament, cytokinesis, histone acetylation, budding cell isotropic bud growth, cell wall organization, biogenesis
EXO84	YBR102C	12	6	16.4	3.8	Spliceosome assembly, exocytosis
CDC6	YJL194W	11	7	14.8	3.6	Pre-replicative complex formation and maintenance
LSM5	YER146W	37	16	17.2	3.6	Nuclear mRNA splicing, via spliceosome, mRNA catabolism, rRNA processing
SKP1	YDR328C	42	17	11.1	3.5	Vacuolar acidification, protein ubiquitination, G1/S transition of mitotic cell cycle, protein complex assembly, G2/M transition of mitotic cell cycle, SCF-dependent proteasomal ubiquitin-dependent protein catabolism
MCM10	YIL150C	7	0	9.9	3.4	Unannotated at time of analysis
SPT5	YML010W	11	6	7.6	3.2	Establishment and/or maintenance of chromatin architecture, RNA elongation from RNA polymerase II promoter, regulation of transcription, DNA-dependent
TAF9	YMR236W	16	9	7.8	3.0	Transcription initiation from RNA polymerase II promoter, establishment and/or maintenance of chromatin architecture, chromatin modification, protein amino-acid acetylation, histone acetylation, G1-specific transcription in mitotic cell cycle
WBP1	YEL002C	9	8	12.2	2.8	N-linked glycosylation, cell cycle

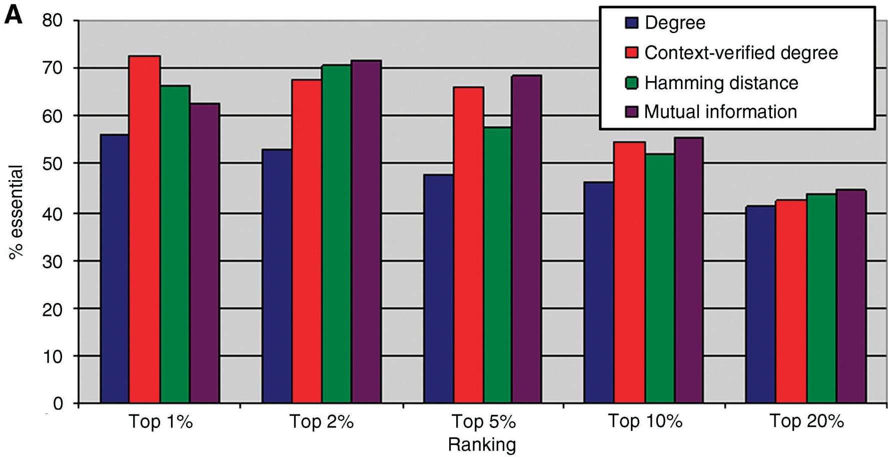
Values provided for context-specific measures, including context degree, context mutual information, and interactive promiscuity, are computed using GO biological process annotations. Measurements based on GO cellular component annotations are included in Supplementary information.

examples (provided in Table I) that present interesting cases of proteins appearing to play a promiscuous role (either directly or in association with other promiscuous proteins), according to the available literature. These self-selected examples are not intended to prove unambiguously the validity of any particular contextual measure, but rather to showcase diverse circumstances by which a particular protein might achieve a high measurement score owing to their cross-contextual associations. All 10 proteins cited are essential for viability. The reader is referred to the supplement for the context measurements of all DIP proteins, based on both GO biological process and GO cellular component annotations. A literature review confirms that these proteins are multifunctional, or involved in complexes that have multiple functional roles. Sec13 is associated with both the nuclear pore complex and the ER. The shuttling of Sec13 between the nucleus and the cytoplasm is believed to play a cross-functional regulatory role (Enninga *et al*, 2003). The high promiscuity of the actin protein Act1 probably reflects the multi-functional role of actin in numerous cellular processes including cell polarization and endocytosis, as well as the diversity of structures formed (cables, contractile rings, and patches). Exo84 has been shown to have dual roles in both the spliceosome and exocytosis (Awasthi *et al*, 2001; Zhang *et al*, 2005). Cdc6 is an origin recognition complex (ORC) constituent whose downregulation is one of several mechanisms to prevent re-initiation (Nguyen *et al*, 2001). Lsm5 is an LSm (Like-Sm) protein whose members form multimeric complexes. Their subunit composition affects both cell location and activity, including mRNA degradation and pre-mRNA intron removal (Zaric *et al*, 2005). Skp1 is associated with multiple functionally diverse protein complexes, including RAVE, CBF3, V-ATPase, and SCFs responsible for regulating a wide variety of cellular processes (Seol *et al*, 2001). The chromatin-bound protein Mcm10 plays a key role in the recruitment of the Mcm2–7 complex to the ORC and multiple other steps in DNA replication initiation (Sawyer *et al*, 2004). Spt5 has been shown to play dual roles in pre-mRNA processing and transcription elongation (Lindstrom *et al*, 2003). Taf9 is a member of TAF (TBP-associated factor) proteins that are constituents of the general transcription factor TFIID and other complexes and highly conserved between human and yeast. Thirteen of the 14 TAFs identified in yeast are essential. TAFs are believed to play a role in the specificity of TFIID. Finally, Wbp1 is a member of the nine-unit oligosaccharyl transferase (OST) complex in yeast. It has been shown that the yeast OST has two functionally distinct isoforms (Schwarz *et al*, 2005), suggesting a broader interpretation for high interactively promiscuous proteins, namely being part of a complex whose variable constituents are themselves active in multiple processes.

Context-sensitive measures are significantly better predictors of essentiality than degree

We identified essential DIP proteins by cross-referencing gene names with a list of 'essential ORFs' from the Saccharomyces Genome Deletion Project (see Materials and methods). We found that proteins ranking highly with respect to various context-sensitive measures are highly enriched in essential proteins, and thus provide a significantly better predictor for protein essentiality than the traditional measure based on node degree alone. For example, we find that context degree is a consistently better predictor for essentiality, which is likely due to the reduction in false-positive edges that otherwise tend to dilute the fraction of essential proteins among the top proteins ranked by node degree. Furthermore, high context-degree nodes are frequently associated with complexes whose component proteins are more uniformly annotated and have inherently high connectivity. Because many of these larger complexes in yeast are essential to cell cycle regulation and mRNA processing, nodes that have the highest context degree are naturally more likely to be essential. We find, for example, that 56% (27 out of 48) of the top 1% of nodes by degree (degree ⩾47) are essential. By comparison, more than 72% of the top 1% of nodes by context degree (context degree ⩾17) are essential. We also note that the average context degree of essential proteins (using GO biological process terms) is 3.9 times greater than the average context degree of non-essential proteins and that although some high context-degree proteins do in fact occur among the set of non-essential proteins, we find that 134/991 (13.5%) of essential proteins have context degree ⩾10, compared to just 60/3750 (1.6%) of non-essential proteins. The key limitation of context degree is that it is not applicable to un-annotated nodes that typically represent 30% or more of the proteins in a PPI network. By contrast, our definitions for interactive promiscuity and context mutual information depend only on the local 'context neighborhood,' that is, the annotations of neighboring proteins, and do not require that the protein itself be annotated. We find that these measures also lead to significant enrichment of essential proteins among top-ranked proteins.

Figure 6 summarizes how various context-sensitive measures for ranking proteins, including interactive promiscuity, perform as a predictor of essentiality. Context mutual information is the best predictor among the top 2% (∼100 proteins), and remains the best predictor as we encompass more proteins (up to 20% of the network). Ultimately, our goal is to identify context-specific criteria that rank proteins such



B

Approximate cutoffs for each measure				
	Degree	Context degree	Average hamming	Total Mut. info
Top 1%	47	17	4.7	28.7
Top 2%	34	13	4.0	19.3
Top 5%	22	9	3.3	8.9
Top 10%	16	5	2.7	4.3
Top 20%	9	2	2.1	1.8
Top 50%	3	0	1.3	0.7

Figure 6 (**A**) Percent of proteins (*N*=4741) that are essential (knockout lethal) for the top-ranked proteins (1–20%), according to various measures including degree, context degree, context mutual information, and interactive promiscuity. The highest ranked proteins (top 1%) using context-verified degree contain the highest proportion of lethal nodes, but this measure is surpassed by the mutual-information-based measure when we include the top 2%. All three measures outperform degree as a predictor of lethality, although, as we encompass larger numbers of proteins, the differences are less pronounced. (**B**) Approximate measure cutoffs for corresponding rank levels.

that the most highly ranked proteins are more likely to be essential. We find that our context-sensitive measures produce a statistically significant increase in the rankings of essential proteins (Wilcoxon rank sum, $\alpha=0.05$, see Materials and methods). Thus, the top-ranked proteins according to each measure are more likely to be essential than the top proteins ranked by node degree alone. It is interesting to note that interactive promiscuity and context mutual information are both associated with the enrichment of essential genes, even though interactive promiscuity is measuring annotation variability whereas context mutual information is measuring correlation among annotations. Although we have not established a conclusive reason for this effect, we have verified that the two measures are not well correlated ($R^2=0.43$). In Supplementary Figure S6, we provide values for all measures averaged across essential versus non-essential genes, and show that context-verified degree and context mutual information exhibit the largest percent increases (close to 300%).

Discussion

It has been argued that scale-free networks are more vulnerable to targeted attacks on hubs than their Erdös-Rényi cousins, because their disruption effectively disconnects the network as a whole. At the same time, such networks are more tolerant of random node removal (Albert *et al*, 2000). This interpretation has important implications for drug target discovery and validation. In one analysis, a correlation between node degree and essentiality was demonstrated in *S. cerevisiae*, where 62% of high-degree proteins (having 15 or more links) were essential compared to 21% overall (Jeong *et al*, 2001). However, more recent yeast knockout data from the Saccharomyces Genome Deletion Project (see Materials and methods) mapped to DIP network nodes yields $P(\text{essential}|\text{degree}\geqslant 15)=0.46$, but with $P(\text{degree}\geqslant 15|\text{essential})=0.24$, suggesting a lack of strong correlation between high-degree nodes, connectivity, and essentiality. We have shown that it is possible to incorporate contextual information to generate more accurate predictors of essentiality. Although it is not the aim of this paper to systematically investigate competing *in silico* techniques for predicting whether or not a gene is essential, we have demonstrated that taking context into account while focusing on degree-type measures does indeed improve our ability to predict essentiality. We expect similar benefits to be garnered when context is applied to other functional problems. Thus, we focused on the added predictive value of features such as context degree versus degree. We suspect that other functional predictors can also be enhanced if appropriately modified to incorporate available context information.

Beyond understanding why certain gene knockouts are lethal, a fundamental goal of systems biology research is to enable one to infer the specific effects of a drug compound by reverse engineering the biological network (Butcher *et al*, 2004). This is based on the hypothesis that if protein B is reachable from protein A by some regulatory or signaling mechanism via protein X, then the perturbation of protein A will affect protein B. However, if the connection from A to B via X is an artifact of aggregate observations, then the conclusions drawn from such networks may not apply. It is also important to note that experimental error may significantly impact network topology (Lin and Zhao, 2005) and any subsequent conclusions.

The limitations of viewing biological networks in purely static terms has been previously demonstrated by showing that changes to network topology resulting from the removal of targeted ‘hubs’ depends very much on whether the hubs are spatial ‘party hubs’ or temporal ‘date hubs’ (Han *et al*, 2004). As noted above, interactively promiscuous and interactively conserved proteins differ in practice from the date and party hub concept owing to the fact that the date/party hub distinction is based on the correlation of expression patterns, whereas we focus on GO process and component annotations as representative of a specific context. In addition, we found that our most interactively promiscuous proteins ($N=190$, $\text{IP}\geqslant 3.5$) contain both party and date hubs in roughly equal numbers (36 party/32 date). We further note that interactive promiscuity measure does not correlate with the average Pearson correlation coefficient used by Han *et al*, to distinguish date and party hubs (see Supplementary Figure S7). It is clear, therefore, that interactive promiscuity and date hubs are capturing unique (and equally valid) characteristics of a protein within a network context.

It has also been observed that existing experimental methods to detect PPIs have intrinsically limited coverage, suggesting that the topology of biological sub-networks cannot be extrapolated to infer the properties of the complete interactome (Han *et al*, 2005). The biological context network model introduced in this paper represents an intermediate level of abstraction for computational modeling of biological networks. Such intermediate level models may play an increasingly important role in systems biology research (Bornholdt, 2005). By contrast, the more complex models involving systems of differential equations, Boolean networks (Akutsu *et al*, 2000; Kauffman *et al*, 2003; Ghim *et al*, 2005; Quayle and Bullock, 2005), or Bayesian networks (Letovsky and Kasif, 2003; Beer and Tavazoie, 2004; Friedman, 2004; Li and Chan, 2004) attempt to capture the full behavior of the network or focus on predicting the regulation of network components under specific conditions (Segal *et al*, 2003) rather than its basic connectivity or neighborhood properties. Of course, any model that introduces a broader range of tunable control parameters is subject to the inherent risk of overfitting. We have found that a generalization of a biological context network (in which cross-contextual interactions are allowed) is equivalent to a Boolean network formalism (Alon *et al*, in preparation). We believe, however, that biological context networks provide a more natural framework for the investigation of certain properties of the network. Such properties include context degree, interactive promiscuity, or whether one protein is ‘reachable’ from another along interaction pathways associated with multiple contexts. In a follow-up paper (Alon *et al*, in preparation), we show, in fact, that the average number of proteins ‘reachable’ from any protein in the network grows linearly in the number of contexts used.

While we maintain the simplicity and analytical advantages of a conventional graphical model, we add one element of complexity to modeling of cellular processes intended to capture PPIs in different functional contexts.

The biological context network formalism provides insight into the statistical topological characteristics of the network within specific contexts, including hubs, dense-sub-graphs, connectivity, and centrality, but quantified with respect to particular contexts. This formalism also suggests an explanation for the emergence of scale-free properties in PPI networks, and offers a measure for the interactive promiscuity of a protein, highlighting those proteins that are either intrinsically multi-functional or are a subunit of a multi-functional complex. This opens up a number of new directions for research. In this paper, we briefly looked at promiscuity of proteins and the conservation of simple protein interaction motifs (in the form of protein pairs) across process-specific and location-specific contexts. We also primarily focused on sub-networks that emerge from applying a single context to the entire interactome, but the formalism naturally captures any systematic application of different context to sub-graphs to allow the study of cross-functional or cross-pathway properties of the network. In particular, we note that using GO annotations as a representative context is inherently limited by the fact that two semantically related but non-ancestral sub-terms are treated as separate and distinct contexts in our analysis. Furthermore, as we pointed out earlier, a context-specific sub-network identifies a set of proteins and interactions that are internally consistent with a specific context, but there is no guarantee that a specific interaction actually occurs in the corresponding process (or location).

Finally, we note that a protein may be active in disparate contexts simultaneously. Explicitly accounting for these multifarious contexts may serve to further refine our notion of the interactive promiscuity and other context-sensitive measures introduced in this paper. Additionally, we may wish to consider interactions occurring in certain context combinations. We have shown, however, that even simple contextual considerations provide a useful perspective on the structure and function of complex interaction networks. Future applications of context-specific functional modules and networks include the modeling of cross-context connectivity and the contextual effects of perturbations on biological function, and enabling improved selection of drug targets by way of more reliable models of toxicity.

Materials and methods

In modeling a PPI network as a context network, we presuppose that two proteins sharing an edge in the original network actually interact only when they share a common context, identified here as the particular GO biological process annotation shared by both proteins. In general, sub-network generation must further accommodate the fact that GO annotations exist as part of a hierarchical directed acyclic graph, meaning therefore that certain node labels or annotations are implicit, although in our analysis, we consider sub-network annotation terms having no descendants (i.e., leaf terms). Context-specific sub-networks are formed by including those proteins that have a particular GO annotation (or a more specific descendant-term annotation) and then carrying over any edges between selected nodes that also occur in the original network. Thus, if we project with respect to the highest level term, GO:0008150—biological process, we effectively reconstitute the entire PPI sub-network of annotated proteins, whereas if we project with respect to some leaf term, say GO: 0006513—protein monoubiquitinization, we will recover only a small set of nodes along with a few edges occurring wherever both incident proteins have this particular annotation explicitly.

Table II *W*-test statistic (Wilcoxon rank sum) and *P*-values for three context-sensitive measures, showing that there is a statistically significant increase in the rankings of essential proteins (versus degree)

Measure	Mean rank	*W*	*P*-value
Context degree)	3160.0	512 452	0.0459
Context mutual information	3075.6	520 558.5	0.0102
Interactive promiscuity	3090.8	512 984	0.0424

PPI network data for *S. cerevisiae* was obtained from the DIP, the DIPs (Xenarios *et al*, 2002) (version 05 June 2005), available online at http://dip.doe-mbi.ucla.edu. This network currently contains 4741 nodes (proteins) connected by 15 429 edges (interactions). We annotated this network by mapping biological process annotations for yeast (Revision 1.1155, 08 June 2005) from the GO consortium http://www.geneontology.org/GO.current.annotations.shtml. The GO consortium provides 12 231 annotations for all yeast gene products. Of these annotations, 6408 (52.3%) covering 970 unique GO terms were mapped to nodes in the DIP network, 1752 (14.3%) are identical annotations with alternative sources of evidence, and 1667 (13.6%) correspond to the GO term 'biological_process unknown' (GO:0000004), and were discarded.

Nodes were annotated as being essential (lethal in knockout experiments) by cross-referencing node names with a list of 'essential ORFs' from data available from the Saccharomyces Genome Deletion Project (http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html.) Of the 1106 essential ORFs provided, 991 (89.6%) map to nodes in the DIP network. Thus, 991 of 4741 (20.9%) DIP nodes were deemed to be essential.

The result that context-sensitive measures cause essential proteins to be ranked higher than proteins ranked by degree was tested for statistical significance using a Wilcoxon rank sum (Mann–Whitney) test ($\alpha=0.05$) applied to all known essential proteins. All 4741 network proteins were first scored by degree, context degree, context mutual information, and interactive promiscuity. We then 'normalized' these scores by converting them to rank values, accounting for ties using the standard procedure of averaging rank scores and applying this result to all tied proteins. The ranks for the 991 essential proteins were then compared between the various measures to determine whether context-sensitive measures produced a statistically significant increase in the rankings of essential proteins versus the rankings based on degree. Our conclusion that each context-sensitive measure produces a statistically significant increase in the rankings of essential proteins is based on the computed *P*-values provided in Table II.

In our simulation of aggregating random (Erdös-Rényi) networks, we constructed a random biological context network consisting of $N=1000$ nodes. Each node contained annotations drawn from a fixed set of 100 unique context labels, subject to the constraint that the number of contexts or labels per protein was distributed according to a power-law distribution with $\gamma=2.00$. We assumed a fixed probability, P, that two proteins in the same contextual state interact. We generated context-specific sub-networks (method described above) for each unique label in random order. Each sub-network was unioned with an aggregate network. We measured the degree distribution of the aggregate network after $L=1,2,3, \ldots, 100$ sub-network aggregations. The resulting degree distribution as averaged across 100 random trials. In each random trial, the number of contextual labels assigned to each protein remains fixed, whereas the context label assignments, edges between proteins, and the generation order of the context-specific sub-networks are randomized.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Noga Alon, Vera Asodi, John Byers, Jim Collins, Tim Gardner, Daniel Hanisch, Sid Redner, and Richard J Roberts for a wide range of helpful suggestions and insights. We also thank our three anonymous reviewers for their extremely helpful comments. This work was supported in part by NSF grants DBI-0239435 and ITR-048715, NHGRI grant #1R33HG002850-01A1, and NIH grant U54 LM008748.

References

Akutsu T, Miyano S, Kuhara S (2000) Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J Comput Biol* **7:** 331–343

Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* **406:** 378–382

Amaral LA, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. *Proc Natl Acad Sci USA* **97:** 11149–11152

Awasthi S, Palmer R, Castro M, Mobarak CD, Ruby SW (2001) New roles for the Snp1 and Exo84 proteins in yeast pre-mRNA splicing. *J Biol Chem* **276:** 31004–31015

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. *Science* **286:** 509–512

Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5:** 101–113

Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. *Cell* **117:** 185–198

Bornholdt S (2005) Systems biology. less is more in modeling large genetic networks. *Science* **310:** 449–451

Butcher EC, Berg EL, Kunkel EJ (2004) Systems biology in drug discovery. *Nat Biotechnol* **22:** 1253–1259

di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol* **23:** 377–383

Enninga J, Levay A, Fontoura BM (2003) Sec13 shuttles between the nucleus and the cytoplasm and stably interacts with Nup96 at the nuclear pore complex. *Mol Cell Biol* **23:** 7271–7284

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. *Science* **303:** 799–805

Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301:** 102–105

Ghim CM, Goh KI, Kahng B (2005) Lethality and synthetic lethality in the genome-wide metabolic network of *Escherichia coli*. *J Theor Biol* **237:** 401–411

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley Jr RL, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM (2003) A protein interaction map of *Drosophila melanogaster*. *Science* **302:** 1727–1736

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430:** 88–93

Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* **23:** 839–844

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32** (Database issue)**:** D258–D261

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* **402** (Suppl 6761)**:** C47–C52

Hoffmann R, Valencia A (2003) Life cycles of successful genes. *Trends Genet* **19:** 79–81

Hood L (2003) Systems biology: integrating technology, biology, and computation. *Mech Ageing Dev* **124:** 9–16

Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. *Science* **306:** 640–643

Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2:** 343–372

Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411:** 41–42

Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci USA* **101:** 2888–2893

Kauffman S, Peterson C, Samuelsson B, Troein C (2003) Random Boolean network models and the yeast transcriptional network. *Proc Natl Acad Sci USA* **100:** 14796–14799

Letovsky S, Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19** (Suppl 1)**:** i197–i204

Li Z, Chan C (2004) Inferring pathways and networks with a Bayesian framework. *FASEB J* **18:** 746–748

Lin N, Zhao H (2005) Are scale-free networks robust to measurement errors? *BMC Bioinformatics* **6:** 119

Lindstrom DL, Squazzo SL, Muster N, Burckin TA, Wachter KC, Emigh CA, McCleery JA, Yates III JR, Hartzog GA (2003) Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. *Mol Cell Biol* **23:** 1368–1378

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431:** 308–312

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* **402:** 83–86

Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U (2004) Superfamilies of evolved and designed networks. *Science* **303:** 1538–1542

Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A (2006) Human protein reference database—2006 update. *Nucleic Acids Res* **34** (Database issue)**:** D411–D414

Nguyen VQ, Co C, Li JJ (2001) Cyclin-dependent kinases prevent DNA re-replication through multiple mechanisms. *Nature* **411:** 1068–1073

Quayle AP, Bullock S (2005) Modelling the evolution of genetic regulatory networks. *J Theor Biol* **238:** 737–753

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, Klitgord N, Simon C, Boxem M, Milstein S, Rosenberg J, Goldberg DS, Zhang

LV, Wong SL, Franklin G, Li S, Albala JS, Lim J, Fraughton C, Llamosas E, Cevik S, Bex C, Lamesch P, Sikorski RS, Vandenhaute J, Zoghbi HY, Smolyar A, Bosak S, Sequerra R, Doucette-Stamm L, Cusick ME, Hill DE, Roth FP, Vidal M (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437:** 1173–1178

Sawyer SL, Cheng IH, Chai W, Tye BK (2004) Mcm10 and Cdc45 cooperate in origin activation in *Saccharomyces cerevisiae*. *J Mol Biol* **340:** 195–202

Schwarz M, Knauer R, Lehle L (2005) Yeast oligosaccharyltransferase consists of two functionally distinct sub-complexes, specified by either the Ost3p or Ost6p subunit. *FEBS Lett* **579:** 6564–6568

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34:** 166–176

Seol JH, Shevchenko A, Shevchenko A, Deshaies RJ (2001) Skp1 forms multiple protein complexes, including RAVE, a regulator of V-ATPase assembly. *Nat Cell Biol* **3:** 384–391

Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci USA* **102:** 1974–1979

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci USA* **100:** 12123–12128

Vazquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, Barabasi AL (2004) The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc Natl Acad Sci USA* **101:** 17940–17945

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, 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

Yook SH, Oltvai ZN, Barabasi AL (2004) Functional and topological characterization of protein interaction networks. *Proteomics* **4:** 928–942

Zaric B, Chami M, Remigy H, Engel A, Ballmer-Hofer K, Winkler FK, Kambach C (2005) Reconstitution of two recombinant LSm protein complexes reveals aspects of their architecture, assembly, and function. *J Biol Chem* **280:** 16066–16075

Zhang X, Zajac A, Zhang J, Wang P, Li M, Murray J, TerBush D, Guo W (2005) The critical role of Exo84p in the organization and polarized localization of the exocyst complex. *J Biol Chem* **280:** 20356–20364

Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S (2002) Computational identification of operons in microbial genomes. *Genome Res* **12:** 1221–1230