Controllability in protein interaction networks

Stefan Wuchty[1,2]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892

Edited by Peter Schuster, University of Vienna, Vienna, Austria, and approved April 8, 2014 (received for review June 12, 2013)

Recently, the focus of network research shifted to network controllability, prompting us to determine proteins that are important for the control of the underlying interaction webs. In particular, we determined minimum dominating sets of proteins (MDSets) in human and yeast protein interaction networks. Such groups of proteins were defined as optimized subsets where each non-MDSet protein can be reached by an interaction from an MDSet protein. Notably, we found that MDSet proteins were enriched with essential, cancer-related, and virus-targeted genes. Their central position allowed MDSet proteins to connect protein complexes and to have a higher impact on network resilience than hub proteins. As for their involvement in regulatory functions, MDSet proteins were enriched with transcription factors and protein kinases and were significantly involved in bottleneck interactions, regulatory links, phosphorylation events, and genetic interactions.

Recently, the focus of modern network research shifted to the determination of nodes that allow the control of an entire network. In particular, Liu et al. (1) introduced a maximum matching approach to predict nodes for the control of various technical, social, and biological networks. Whereas their approach only applied to directed networks, Nacher and Akutsu (2) suggested an equivalent optimization procedure to determine minimum dominating sets (MDSets) of nodes that play an important role for the control of undirected networks. An intriguing question, however, remains if such nodes carry important functional characteristics. Generally, the importance of a protein in an interaction network is frequently considered a question of the number of interactions a given protein is involved in. For instance, the so-called centrality–lethality rule was first suggested by Jeong et al. (3) and Yu et al. (4), stating that highly connected proteins tend to be essential. Furthermore, such hubs are also involved in a rising number of protein complexes (5), suggesting that their essentiality is a consequence of their complex involvement (6, 7). In humans, human viruses and parasites target certain proteins to seize control of a host cell (8, 9) whereas such proteins play a decisive role in different cancer types (10, 11). Therefore, we wondered whether protein sets that are predicted to be important for the control of a protein interaction network would carry such biological significance as well. In other words, we expected that minimum dominating sets of proteins were enriched with, for example, disease or essential genes. Focusing on the currently best investigated interactomes we determined MDSets in human and yeast. Such sets are defined as finite subsets of proteins from where each remaining protein can be immediately reached by one interaction. Strongly suggesting that such well-defined protein groups have significance, MDSet proteins were indeed enriched with essential, cancer-related and virus-targeted genes. Furthermore, we found that MDSet proteins were preferably placed in central network positions, enabling MDSet proteins to connect protein complexes and significantly appear in bottleneck interactions, regulatory and phosphorylation events, and genetic interactions.

Results

In a protein interaction network, an MDSet is defined as an optimized subset of proteins from where each remaining (i.e., non-MDSet) protein can be reached by one interaction. Therefore, each non-MDSet protein is connected to at least one MDSet protein (Fig. 1*A*). In protein interactions of *Homo sapiens* and *Saccharomyces cerevisiae* of the High-quality INTeractomes (HINT) database (12) we determined corresponding minimum dominating sets by solving an integer-based linear programming problem (*Methods*). Although we considered their combination, we separately accounted for binary and cocomplex interactions in each organism as well (*Methods*). The table in Fig. 1*B* indicates that the corresponding MDSets of human and yeast interaction networks involved fewer than 20% of all proteins. Compared with the mean degree of 6.7 in the combined human interaction network, the mean degree of MDSet proteins increased to 17.1. In the combined yeast network, each protein was on average involved in 10.0 interactions and the mean degree of MDSet proteins rose to 23.8. Such trends also applied to binary and cocomplex interaction sets of both organisms (Fig. 1*B*). Whereas the degree distributions of all proteins in interaction networks are generally characterized by fat tails (3, 13), the degrees of MDSet proteins showed the same distribution (Fig. S1*A*). To determine the enrichment of MDSet proteins as a function of their number of interactions we grouped proteins according to their degree in bins of logarithmic size. In each group we compared the protein's frequency distributions in the combined interaction networks and the corresponding MDSets (*Methods*). Fig. 1*C* clearly demonstrates that MDSets were mostly enriched with proteins that roughly had more than 10 interactions. Fig. S1*B* shows similar results in the binary and cocomplex interaction sets of both organisms. Indicating a protein's central role in an interaction network, we calculated a protein's betweenness centrality. Fig. S2*A* shows that the frequency distributions of betweenness centralities of MDSet proteins in all interaction sets have fat tails. In Fig. S2*B*, we grouped proteins in bins of logarithmic size and compared the protein's frequency distributions in the underlying interaction networks and their corresponding MDSets (*Methods*). Specifically, we observed that

Significance

In human and yeast protein interaction datasets we determined minimum dominating sets (MDSets), proteins that play a role in the control of the underlying interaction webs. Such proteins are defined as optimized subsets from where each remaining protein can be immediately reached. Notably, MDSet proteins were enriched with cancer-related and virus-targeted genes. Furthermore, MDSet proteins have a higher impact on network resilience than hub proteins. Indicating their relevance for the controllability of biological networks, we also found a strong involvement in bottleneck interactions, regulatory and phosphorylation events as well as genetic interactions.

Author contributions: S.W. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The author declares no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]Present address: Department of Computer Science and Center of Computational Sciences, University of Miami, Coral Gables, FL 33146.

[2]E-mail: wuchtys@cs.miami.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1311231111/-/DCSupplemental.

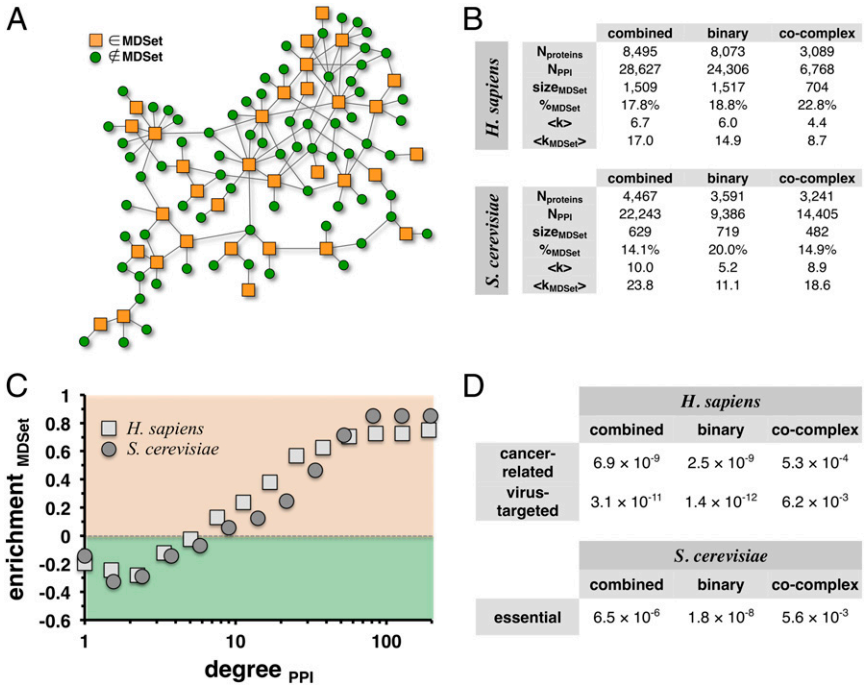

Fig. 1. Definition and characteristics of minimum dominating sets. (*A*) In a toy network we illustrate the concept of an MDSet. Specifically, an MDSet is defined as an optimized subset of nodes (orange squares) from where each remaining (i.e., non-MDSet) node (green circles) can be immediately reached by one step. As a consequence, each non-MDSet protein is linked to at least one MDSet protein. (*B*) In the table we present statistics of protein interaction networks and their corresponding MDSets in human and yeast. In particular, we accounted for binary and cocomplex interactions as well as combined interaction datasets. (*C*) We grouped proteins in logarithmic bins according to their number of interaction partners in the combined networks. In each bin, we separately calculated the frequency of proteins in the full networks and MDSets. In comparison, MDSet proteins mostly appeared enriched in groups of proteins with roughly more than 10 interactions. (*D*) By applying a Fisher's exact test, we found that cancer-related genes and proteins that are targeted by human viruses are significantly enriched in human MDSets. Similarly, essential genes are significantly present in the corresponding yeast MDSets.

proteins with increasing betweenness predominantly appeared in the corresponding MDSets of all different interaction datasets of human and yeast.

To account for false positives in the underlying interaction networks, we simulated their presence by randomly deleting a certain fraction of interactions. To investigate the robustness of MDSets toward such perturbations we evaluated the overlap of MDSets in the actual and the randomized networks with a Jaccard index. Fig. S3 shows that the overlap between the MDSets of the actual and randomized networks decreased but seemed to steady with an increasing rate of false positives. Notably, such trends applied to yeast and human networks of all interaction types.

To indicate the biological significance of MDSet proteins in a human interaction network we hypothesized that such sets may be significantly enriched with proteins that govern diseases. In particular, we collected 496 oncogenes and 876 tumor suppressor genes from the CancerGenes database, totaling 1,370 cancer-related genes (14). Applying Fisher's exact test, we found that cancer-related genes were indeed significantly enriched in the MDSet of the combined human interaction network ($P = 6.9 \times 10^{-9}$). In human binary and cocomplex interaction data we obtained statistically significant enrichments of cancer genes as well ($P < 10^{-3}$, Fig. 1*D*). Analogously, we used 770 human proteins that were involved in 4,474 interactions with proteins of various human viruses from the Molecular INTeraction (MINT) database (15). In all human interaction sets, targeted proteins were significantly enriched in the corresponding MDSets ($P < 0.01$, Fig. 1*D*) with the binary network showing the strongest signal ($P = 1.4 \times 10^{-12}$). To show the importance of MDSet proteins in yeast interaction networks we determined the enrichment of 1,110 essential genes in *S. cerevisiae* that were compiled from the Database of Essential Genes (DEG) (16). We observed that essential genes were significantly enriched in the MDSets of all yeast interactions datasets ($P < 0.01$, Fig. 1*D*), with the binary interaction network showing the strongest enrichment ($P = 1.8 \times 10^{-8}$).

To compare the enrichment levels of disease and essential genes we created sets of most connected proteins that matched the size of the corresponding MDSets in each interactions dataset. In Fig. S4*A*, we calculated the enrichment of cancer-related genes and proteins that were targeted by human viruses. We found that sets of most connected proteins showed much stronger enrichments of such disease genes, results that applied for essential genes in yeast as well. Analogously, we obtained similar results with sets of most central proteins in human and yeast interaction sets (Fig. S4*B*).

Protein complexes provide a different functional level of molecular organization, suggesting that MDSet proteins may appear in more complexes. Using data from the CORUM (Comprehensive Resource of Mammalian protein complexes) database, we collected 1,843 protein complexes in *H. sapiens* (17) and 409 protein complexes in *S. cerevisiae* from the CYC2008 database (18). Fig. S5 clearly indicates that MDSet proteins in the combined human interaction network appeared in significantly more protein complexes than non-MDSet proteins ($P = 2.6 \times 10^{-8}$, Wilcoxon test), results that also held for yeast protein complexes ($P = 4.1 \times 10^{-5}$). Furthermore, we determined the complex participation coefficient P_i of each protein i (19). P_i tends toward 1 if i interacts with proteins in the same protein complexes, and vice versa. Because MDSets are most central to reach all other proteins in a network, we hypothesized that MDSet proteins preferably connect many different protein complexes through their interactions. Therefore, we expected that MDSet proteins have lower complex participation coefficients. Comparing the

coefficient distributions of (non)MDSet proteins in the combined yeast interaction network in Fig. S6, we indeed found that MDSet proteins provided lower values than non-MDSet proteins ($P = 3.3 \times 10^{-17}$, Wilcoxon test), a result that also applied to complexes in the combined human interaction network ($P = 2.4 \times 10^{-38}$).

To measure a protein's impact on an interaction network's resilience, we performed a robustness analysis. Using the combined yeast network, we sorted all 629 MDSet proteins according to their degree. To compare, we created sets of equal size of the most connected yeast proteins as well as randomly picked proteins. Starting with the most connected protein we gradually deleted proteins and calculated the number of connected components after each deletion step. Successive deletion of MDSet proteins had a higher impact by producing more connected components and removing fewer interactions than their hub and random counterparts (Fig. 2*A*). In Fig. S7, we observed that such perturbations provided similar results in the combined human network.

Whereas single MDSet proteins generally showed strong enrichments, we expected that similar signals emerge from topological and functional interactions between MDSet proteins. Specifically, we focused on bottleneck interactions, defined as the top 10% of interactions with the highest edge betweenness (20). In both the combined human and yeast networks we counted the number of bottleneck interactions that involved pairs of (non)MDSet proteins. As a null model, we randomly sampled MDSet proteins 10,000 times and expected to find similar numbers if their placement was a random process. After determining the corresponding enrichment/depletion of such bottleneck interactions we clearly observed that bottleneck interactions were significantly enriched between MDSet proteins whereas the opposite held for pairs of non-MDSet proteins ($P < 10^{-4}$, Fig. 2*B*).

Genetic interactions can reveal important functional relationships between genes and pathways (21), suggesting that genetic interactions may be overrepresented between MDSet proteins. After collecting 108,899 genetic interactions between 5,364 genes in *S. cerevisiae* from the Biological General Repository for Interaction Datasets (BioGRID) database (22), we counted genetic interactions between (non)MDSet proteins. Randomly sampling MDSets 10,000 times we clearly observed that genetic interactions are significantly enriched when at least one protein participated in the MDSet (Fig. 2*C*). In turn, the opposite held for genetic interactions between non-MDSet proteins ($P < 10^{-4}$).

Assuming that MDSets may significantly contribute to control processes we hypothesized that transcription factors and their target genes may significantly appear in MDSets. Specifically, we used 95,722 regulatory interactions between 209 human transcription factors and 8,910 target genes from the TRANScription FACtor (TRANSFAC) database (23, 24). Furthermore, we assumed that the same logic applies to phosphorylation events and collected 5,462 human phosphorylation events between 207 kinases and 1,661 from the networKIN database (25, 26). Applying Fisher's exact test we found that transcription factors ($P = 2.7 \times 10^{-4}$) and kinases ($P = 3.4 \times 10^{-12}$) were significantly enriched in the MDSet of the combined human interaction network. Additionally, we counted how often a pair of transcription factors and a given target gene appeared between (non)MDSet proteins. Specifically, we observed that regulatory interactions and phosphorylation events were significantly enriched when corresponding transcription factors and kinases were involved in the MDSet ($P < 10^{-4}$, Fig. 2*D*). In turn, interactions

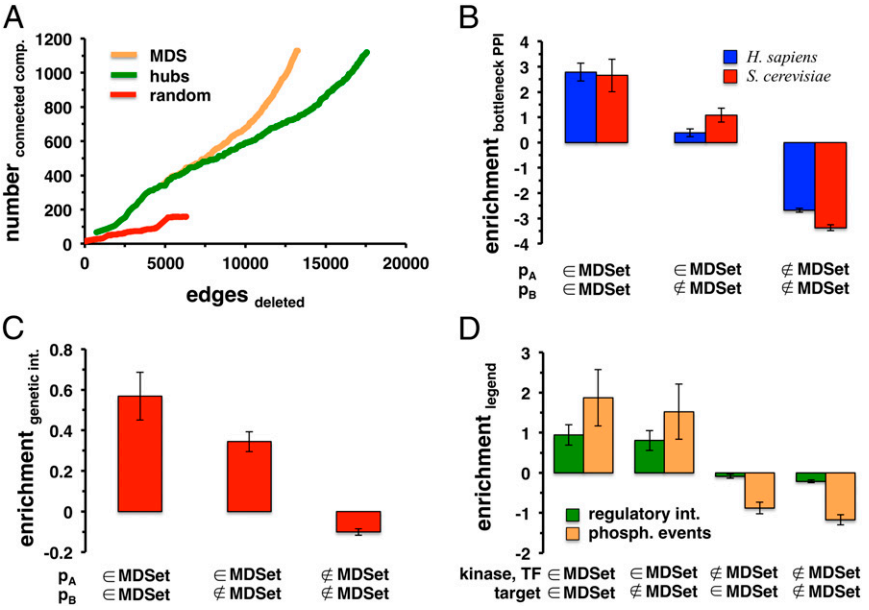

Fig. 2. Enrichment of topological and functional entities in MDSets. In *A* we sorted all MDSet proteins in the combined yeast interaction network according to their degree. To provide an equivalent set of equal size we collected and sorted the highest connected hub proteins. Furthermore, we randomly sampled a set of yeast proteins of the same size. Starting with the most connected protein, we gradually deleted proteins and calculated the number of connected components in the altered network. In comparison, the deletion of MDSet proteins had a higher impact on the resilience of the underlying networks than hubs alone. (*B*) Defined as the top 10% of interactions with the highest edge betweenness, we determined a set of bottleneck interactions in the combined human and yeast networks, respectively, and counted their occurrence between (non)MDSet proteins (p_A and p_B). Randomly sampling MDSets 10,000 times, we observed that bottleneck interactions were significantly enriched between MDSet proteins and depleted between non-MDSet proteins ($P < 10^{-4}$). In *C* we determined the enrichment of genetic interactions between yeast (non)MDSet proteins (p_A and p_B) in the combined network. Randomly sampling MDSet proteins 10,000 times we clearly observed that genetic interactions preferably appeared between MDSet proteins, whereas the opposite applied for non-MDSet proteins ($P < 10^{-4}$). (*D*) Human phosphorylation events and transcriptional, regulatory interactions were significantly enriched when at least kinases or transcription factors were involved in the MDSet of the combined human network ($P < 10^{-4}$).

between a transcription factor and a target gene seemed generally depleted when both were not involved in the MDSet ($P < 10^{-4}$). In yeast, we used 48,082 regulatory interactions between 183 yeast transcription factors and 6,403 genes from the Yeast Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT) database (27). Furthermore, we obtained 3,466 experimentally determined interactions between 80 kinases and 1,172 substrates from (28), allowing us to find similar, albeit less significant, enrichment patterns (Fig. S8).

Discussion

Here, we determined minimum dominating sets of proteins in interaction networks that were defined as the smallest group of strategically placed proteins from where each remaining protein (i.e., non-MDSet protein) can be immediately reached through an interaction. As a consequence each non-MDSet protein therefore interacts with at least one MDSet protein. Although we observed that MDSet proteins are enriched among highly connected proteins, a large degree is not necessarily a criterion that qualifies a protein to participate in the MDSet. Notably, we observed that the degree distributions of MDSet proteins have fat tails, indicating that the majority of MDSet proteins have a small number of interaction partners, and vice versa. Such a characteristic is quite different compared with hubs that are widely considered the topologically and functionally most important proteins in an interaction network. In particular, the only criterion to consider a protein a hub is a preferably large number of interaction partners. Furthermore, the definition of hubs depends on an arbitrarily set threshold that only rigidly accounts for the local vicinity of a node. In turn, the way to determine MDSets considers the whole network, providing an optimal smallest set of strategically placed proteins, a procedure that does not need any arbitrary parameters. Still, MDSets manage to capture a considerable amount of highly connected proteins. Furthermore, MDSet proteins preferably appeared among proteins of high betweenness (i.e., bottleneck nodes), an observation that translated into bottleneck interactions as well. The direct comparison of MDSets with sets of protein hubs is a difficult undertaking, given that no generally applicable threshold or method for the detection of hubs actually exists. However, we generated sets of the most connected proteins that match the size of MDSets as an approximation. To directly compare the topological impact of MDSet proteins and hubs we sorted proteins according to their degree and successively deleted proteins from the underlying network. Notably, the deletion of MDSet proteins had a higher disruptive effect on the underlying network than hub proteins, demonstrating the topological relevance of MDSets.

On a different, more biologically relevant, level of network organization, we found that their strategic placement allowed MDSet proteins to participate in significantly more protein complexes than non-MDSet proteins. Furthermore, their interactions enabled MDSet proteins to reach more proteins in other complexes than non-MDS proteins. Whereas such observations indicate that MDSet proteins reach other proteins effectively, the question remains whether such characteristics translate into a governing role in the underlying networks. Indeed, we found that cancer-related genes and proteins that are targeted by human viruses are enriched with MDSet proteins. Onco- and tumorsuppressor genes play a fundamental causal role for the emergence of tumors whereas proteins that are targeted by viruses form a host–pathogen interface, allowing viruses to interfere with functions in the underlying host cell. Therefore, MDSet proteins may be important for the dissemination of causal information because their central placement provides a topological basis to reach all other proteins efficiently. In a similar vein, the central placement of MDSet proteins may complement functional interactions that exert biological control. In particular, transcription factors govern the expression of their underlying target genes, whereas kinases control the level of phosphorylation of their substrates as an effective means to process biological signals. Genetic interactions between genes indicate potential synergies when mutations in two genes may produce an unexpected phenotype given each mutation's individual effects. Notably, genetic interactions preferably appeared when the interacting proteins were involved in MDSets. The strong involvement of MDSet proteins seems plausible, assuming that a genetic interaction may provide control of compensatory pathways or protein complexes. Considering expression and phosphorylation events, we obtained strongest enrichment signals when both the controlling (i.e., transcription factors, kinases) and controlled entity (i.e., target genes, substrates) occurred in the MDSets. In turn, such interactions seemed most diluted when both transcription factors/kinases and targets/substrates did not participate in the underlying MDSet. Such observations suggest that the topological characteristics of MDSets may be tapped for the collection and dissemination of biological information by transcription factors and kinases. Given that MDSet proteins connect to each remaining protein in the underlying networks by at most one step a transcription factor or kinase that participates in the MDSet may have an advantage to efficiently receive signals through corresponding interactions. In turn, a signal that is mediated by the expression levels of a target gene or the phosphorylation of substrate may have stronger efficacy when distributed through the interactions of an MDSet protein. Therefore, MDSets may be considered a complement that allows transcription and phosphorylation events to efficiently control biological processes.

Methods

Protein–Protein, Regulatory, and Phosphorylation Interactions. We used a total of 28,627 high-quality protein interactions between 8,495 human proteins as well as 22,243 interactions between 4,467 yeast proteins from the HINT database (12). Accounting for methods that allow the detection of binary and cocomplex interactions (29) we obtained 27,254 binary interactions between 8,233 proteins and 7,692 cocomplex interactions between 3,188 proteins in human. As for yeast, we collected 11,435 binary interactions between 3,653 proteins and 16,294 cocomplex interactions between 3,380 proteins. Checking the interaction's quality, Fig. S9 shows that the majority of binary interactions were confirmed by more than one publication. Cocomplex interactions were only accounted for when they were reported in at least two publications.

We collected 95,722 links between 209 human transcription factor and 8,910 human genes from the TRANSFAC (24) database as provided by mSigDB (23). As for regulatory interactions in yeast we used 48,082 regulatory interactions between 183 transcription factors and 6,403 genes from the YEASTRACT database (27). Specifically, such regulatory interactions were indicated if a binding site of given transcription factor appeared in the promoter of the underlying genes.

As for phosphorylation events in human we obtained 5,864 interactions between 63 kinases and 1,452 human proteins from the networKIN database (25, 26). Such links represent a kinase specific phosphorylation site in a given protein. Furthermore, we collected 3,466 experimentally determined phosphorylation events between 80 kinases and 1,172 substrates in yeast (28).

Determination of a Minimum Dominating Set. A set $S \subseteq V$ of nodes in a network $G = (V, E)$ is defined as an MDSet if every node $v \in V$ is either an element of S or adjacent to an element of S. In other words, an MDSet is an optimized subset of nodes from where each remaining node can be immediately reached by one interaction (Fig. 1*A*). Specifically, we modeled and solved a binary integer-programming problem where each protein $v \in V$ that participates in interactions E in a protein interaction network $G = (V, E)$ is assigned a binary variable x_v. If v is an element of the MDSet we defined $x_v = 1$, and 0 otherwise. We modeled the determination of an MDSet as $\min \sum_{v \in V} x_v$, subject to the constraint $x_v + \sum_{w \in \Gamma(v)} x_w \geq 1$, where $\Gamma(v)$ was the set of interaction partners of protein v. Because the domination problem in graphs is NP-complete no algorithm necessarily exists that allows the determination of a minimum dominating set in arbitrary graphs in polynomial time (30). Specifically, we used a branch-and-bound algorithm (31)

(see *SI Methods* and Fig. S10 for more details) as implemented by library lpSolve of the R programming language to solve our binary integer-programming problem.

Essential Genes in *S. cerevisiae*. We used 1,110 essential genes from the DEG database, which collects data about essential genes from the literature (16).

Disease Genes in *H. sapiens*. We collected 496 oncogenes and 876 tumor suppressor genes from the CancerGenes database (14), which collects such information from the literature. Furthermore, we considered 4,474 interactions between proteins of various human viruses and 770 human proteins that the MINT database collected from the literature (15).

Protein Complexes. We used 1,843 protein complexes in *H. sapiens* from the CORUM database (17) and 409 protein complexes in *S. cerevisiae* from the CYC2008 database (18). Both databases collect information about experimentally determined protein complexes from the literature.

Protein Complex Participation Coefficient. For each protein that is involved in at least one protein complex, we defined the protein complex participation coefficient of a protein i as $P_i = \sum_{s=1}^{N} (n_{i,s}/\sum_{s=1}^{n} n_{i,s})^2$, where $n_{i,s}$ is the number of links that protein i had to proteins in complex s out of N total complexes. If a protein predominantly interacted with partners of the same complex, P tended to *1*, and vice versa (32).

Enrichment Analysis. Using a protein interaction network, we grouped proteins according to their degrees or betweenness centrality in bins of logarithmically increasing size. In each group i we determined the corresponding frequency of proteins with a certain characteristic A, $f_{A,i} = N_{A,i}/\sum_i N_{A,i}$. Analogously, we calculated the corresponding frequency of proteins with characteristic A that appeared in a minimum dominating set (MDSet), $f_{A,i}^{MDSet} = N_{A,i}^{MDSet}/\sum_i N_{A,i}^{MDSet}$. Finally, we defined the enrichment of proteins with characteristic A that appear in the MDSet in bin i as $E_{A,i}^{MDSet} = \lg(f_{A,i}^{MDSet}/f_{A,i})$. Therefore, $E_{A,i}^{MDSet} > 0$ points to an enrichment of feature A, and vice versa.

As for the enrichment of genetic interactions, regulatory interactions, or bottleneck interactions between (non)MDSet protein pairs, we counted the number of pairs that are connected by such links, N_A. Randomly sampling minimum dominating sets, we analogously counted the corresponding random number, $N_{r,A}$, and defined the enrichment of such interactions as $E_A = \lg(N_A/N_{r,A})$.

Betweenness Centrality. As a global measure of its centrality, we calculated an edges betweenness, indicating an interactions appearance in shortest paths through the whole network. In particular, we defined betweenness centrality c_B of an edge e as $c_B(e) = \sum_{s \neq t \in V} \sigma_{st}(e)/\sigma_{st}$, where σ_{st} was the number of shortest paths between proteins s and t and σ_{st} (e) was the number of shortest paths running through e. Analogously, we determined the betweenness centrality of node v as $c_B(v) = \sum_{s \neq t \neq v \in V} \sigma_{st}(v)/\sigma_{st}$. Furthermore, we normalized a node v's centrality by $(N-1)(N-2)/2$, where N is the total number of nodes in the network.

ACKNOWLEDGMENTS. We thank A.-L. Barabási, Peter Uetz, and Sawsan Khouri for fruitful discussions. This work was supported by the National Institutes of Health/Department of Health and Human Services (Intramural Research program of the National Library of Medicine) as well as start-up funds from the Department of Computer Science at the University of Miami.

1. Liu YY, Slotine JJ, Barabási AL (2011) Controllability of complex networks. *Nature* 473(7346):167–173.
2. Nacher J, Akutsu T (2012) Dominating scale-free networks with variable scaling exponent: Heterogeneous networks are not difficult to control. *New J Phys* 14(7): 073005–073028.
3. Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411(6833):41–42.
4. Yu H, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* 322(5898):104–110.
5. Wuchty S, Almaas E (2005) Peeling the yeast protein network. *Proteomics* 5(2): 444–449.
6. Zotenko E, Mestre J, O'Leary DP, Przytycka TM (2008) Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. *PLOS Comput Biol* 4(8):e1000140.
7. Song J, Singh M (2013) From hub proteins to hub modules: The relationship between essentiality and centrality in the yeast interactome at different scales of organization. *PLOS Comput Biol* 9(2):e1002910.
8. Batada NN, Hurst LD, Tyers M (2006) Evolutionary and physiological importance of hub proteins. *PLOS Comput Biol* 2(7):e88.
9. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* 296(5568):750–752.
10. Kar G, Gursoy A, Keskin O (2009) Human cancer protein-protein interaction network: A structural perspective. *PLOS Comput Biol* 5(12):e1000601.
11. Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22(18):2291–2297.
12. Das J, Yu H (2012) HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* 6:92.
13. Lehner B, Fraser AG (2004) A first-draft human protein-interaction map. *Genome Biol* 5(9):R63.
14. Higgins ME, Claremont M, Major JE, Sander C, Lash AE (2007) CancerGenes: A gene selection resource for cancer genome projects. *Nucleic Acids Res* 35(Database issue): D721–D726.
15. Zanzoni A, et al. (2002) MINT: A Molecular INTeraction database. *FEBS Lett* 513(1): 135–140.
16. Zhang R, Ou HY, Zhang CT (2004) DEG: A database of essential genes. *Nucleic Acids Res* 32(Database issue):D271–D272.
17. Ruepp A, et al. (2010) CORUM: The comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Res* 38(Database issue):D497–D501.
18. Pu S, Wong J, Turner B, Cho E, Wodak SJ (2009) Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res* 37(3):825–831.
19. Wuchty S, Siwo GH, Ferdig MT (2011) Shared molecular strategies of the malaria parasite P. falciparum and the human virus HIV-1. *Mol Cell Proteomics* 10(10):M111 009035.
20. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M (2007) The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLOS Comput Biol* 3(4):e59.
21. Costanzo M, et al. (2010) The genetic landscape of a cell. *Science* 327(5964):425–431.
22. Stark C, et al. (2006) BioGRID: A general repository for interaction datasets. *Nucleic Acids Res* 34(Database issue):D535–D539.
23. Subramanian A, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102(43):15545–15550.
24. Matys V, et al. (2006) TRANSFAC and its module TRANSCompel: Transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34(Database issue):D108–D110.
25. Linding R, et al. (2007) Systematic discovery of in vivo phosphorylation networks. *Cell* 129(7):1415–1426.
26. Linding R, et al. (2008) NetworKIN: A resource for exploring cellular phosphorylation networks. *Nucleic Acids Res* 36(Database issue):D695–D699.
27. Abdulrehman D, et al. (2011) YEASTRACT: Providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res* 39(Database issue):D136–D140.
28. Ptacek J, et al. (2005) Global analysis of protein phosphorylation in yeast. *Nature* 438(7068):679–684.
29. De Las Rivas J, Fontanillo C (2010) Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLOS Comput Biol* 6(6): e1000807.
30. Haynes TW, Hedetniemi ST, Slater PJ *Fundamentals of Domination in Graphs*. Pure Applied Mathematics (Marcel Dekker, New York).
31. Land AH, Doig AG (1960) An automatic method of solving discrete programming-problems. *Econometrica* 28(3):497–520.
32. Wuchty S, Siwo G, Ferdig MT (2010) Viral organization of human proteins. *PLoS ONE* 5(8):e11796.

Supporting Information

Wuchty 10.1073/pnas.1311231111

SI Methods

Formulation of the Binary Integer Programming Problem. In general, binary integer programming revolves around the problem of finding a binary vector x that optimizes a linear function $f^T x$ subject to linear constraints. For example, a binary integer programming problem may be $min\ f^T x$ such that $Ax \geq b$, where x only can assume binary values. In particular, A is a matrix that holds the coefficients of the constraining variables, and b holds the boundaries of the constraints.

In Fig. S10*A* we consider a simple network whose minimum dominating set (MDSet) we want to determine. Each node v_i can either take value $x_i = 1$ if v_i is part of the MDSet or 0 otherwise. In Fig. S10*B*, we formulate the underlying binary integer programming problem for our toy network. In particular, we want to minimize the number of nodes in the MDSet by min $(x_1 + x_2 + x_3 + x_4)$. Because we defined the MDSet as the minimum subset of nodes that allow us to reach each remaining non-MDSet node by one step (i.e., non-MDSet nodes are in the vicinity of an MDSet node), we need to make sure that the sum of x's of a given node v_i and its neighbors is at least 1. Therefore, our constraints can be formulated for each node v_i $(i = 1,2,3,4)$ as $x_i + \sum_{j \in \Gamma(i)} x_j \geq 1$, where $\Gamma(i)$ are the neighbors of v_i in the toy network. As a consequence, the matrix notation of the constrains is $Ax \geq 1$ where A is the adjacency matrix of the underlying toy network, where $A_{ii} = 1$ as well as $A_{ij} = 1$ if there exists an edge between nodes v_i and v_j and 0 otherwise.

Solving a Binary Integer Programming Problem with a Branch-and-Bound Algorithm. lpSolve of the R package uses a linear programming (LP)-based branch-and-bound algorithm (1) to solve such a binary integer programming problem. The algorithm searches for an optimal solution to the binary integer programming problem by solving a series of LP-relaxation problems, in which the binary integer requirement on the variables is replaced by the weaker constraint $0 \leq x_i \leq 1$. Briefly, the algorithm (*i*) searches for a binary integer feasible solution, (*ii*) updates the best binary integer feasible solution so far as the search tree grows, and (*iii*) verifies that no better integer feasible solution is possible by solving a series of linear programming problems.

In more detail, the algorithm creates a search tree by repeatedly adding constraints of the problem (Fig. S10*C*). This step is called "branching" where the algorithm chooses a variable x_j that has not been set to an integer value yet. Specifically, the algorithm adds the constraint $x_j = 0$ to form one branch and the constraint $x_j = 1$ to form the other branch. In our example the algorithm starts out to add the constraint revolving around node v_1 to the tree by adding a node where x_1 is set to 0 and another node, where $x_1 = 1$. As a consequence such branching steps generate a binary tree. In general, however, the order of the variables going down the levels in the tree is not necessarily the usual order of their subscripts.

At each node, the algorithm solves an LP-relaxation problem based on the constraints that were used up to this node. In our example, let us consider the node where $x_1 = 0$. Here, x_1 is set to 0, but the remaining variables x_2, x_3, and x_4 are still free to take on either value. In other words, for each variable x_2, x_3, and x_4 we consider a relaxed constraint in the interval [0,1]. In the node where $x_1 = 0$, we, for instance, assume that $x_2 = 0.5$, $x_3 = 0.9$, and $x_4 = 0.8$, resulting in $Z = 2.2$. Such a feasible solution (i.e., no violation of the constraints) to the LP-relaxation problem provides a lower bound for the binary integer programming problem. Inevitably, we will end up with a feasible solution where all variables are set to an integer value (i.e., an integer solution). In this case such a binary integer vector would provide an upper bound or the best current integer solution Z^* for the binary integer programming problem. As a start, when no actual solution is yet available we choose a default of $Z^* = \infty$.

Depending on the outcome of this evaluation step, the algorithm will decide either to continue branching (i.e., add another constraint) or to move to another node. In particular, we consider three possibilities in our toy model:

i) If the LP-relaxation problem at the current node is infeasible or Z is greater than the corresponding value of the best current integer solution Z^*, the algorithm will not search any branches below that node. The algorithm then moves to a new node according to the underlying, implemented search strategy. In our example, the algorithm moved to the other branch where $x_2 = 1$, after we found that the solution to the subproblem with $x_2 = 0$ was infeasible.

ii) If a (new) feasible integer solution with a lower Z value than that of the best current integer solution is found ($Z < Z^*$), then the algorithm keeps this solution as the new best current integer solution and moves on to the next node. In our toy example we found a best current integer solution (and the ultimately optimal solution) where $x_2 = 1$, whereas all other variables were 0, and $Z = Z^* = 1$. Our algorithm continued searching at the last branching step where we left off ($x_3 = 1$). Although the solution was feasible, the corresponding value $Z = 1.9$ was larger than $Z^* = 1$.

iii) If the solution of the LP-relaxation problem is feasible but not integer (i.e., a possible final solution) and the Z of the LP-relaxation problem is less than the best current integer solution $Z < Z^*$, we start a new branching step. Following the path to the optimal solution, we observed a series of branching steps in our toy example, illustrating this step.

Following this algorithmic outline, we end up with an optimal solution of $Z^* = 1$ and $x_1 = 1$, whereas $x_{1,3,4} = 0$, suggesting that the MDSet of our toy network consists of node v_2 only.

1. Land AH, Doig AG (1960) An automatic method of solving discrete programming-problems. *Econometrica* 28(3):497–520.

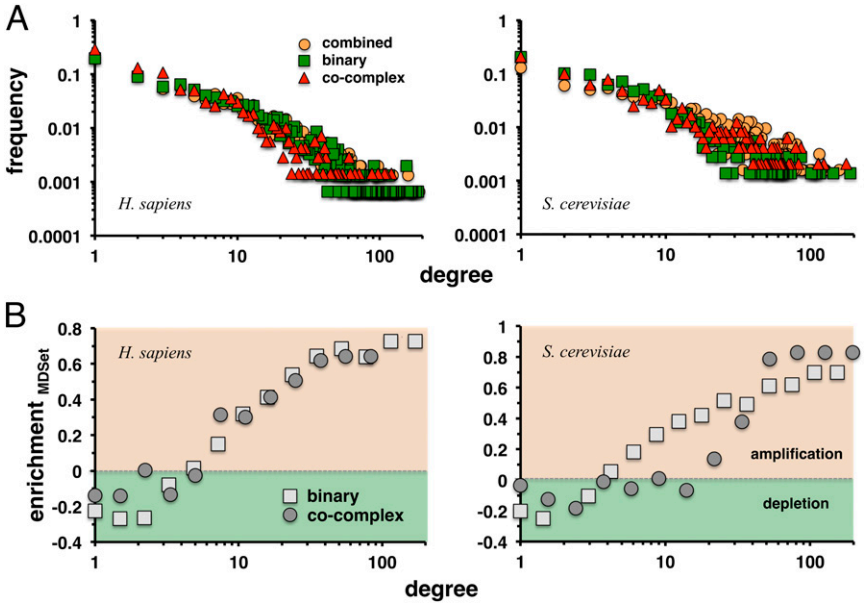

Fig. S1. Degree distributions of MDSet proteins. The degree distributions of proteins that are involved in the MDSet of the combined, binary, and cocomplex interactions in (*A*) human and yeast have fat tails. (*B*) Compared with the log-binned frequency distributions of degrees in the binary and cocomplex interaction datasets, we observed that MDSet proteins are mostly enriched among highly connected proteins in human and yeast.

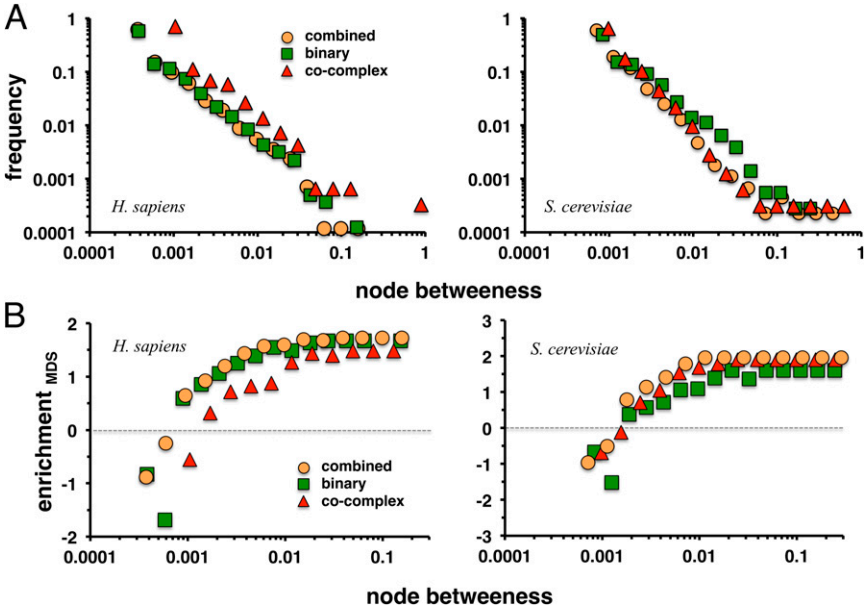

Fig. S2. Betweenness centrality distributions of MDSet proteins. (*A*) The frequency distributions of protein's betweenness centrality that are involved in the MDSet of the combined, binary, and cocomplex interactions in human and yeast have fat tails. (*B*) Determining node-specific betweenness centrality of all proteins in the binary and cocomplex interaction datasets, we observed that MDSet proteins are mostly enriched among proteins with high betweenness in both human and yeast.

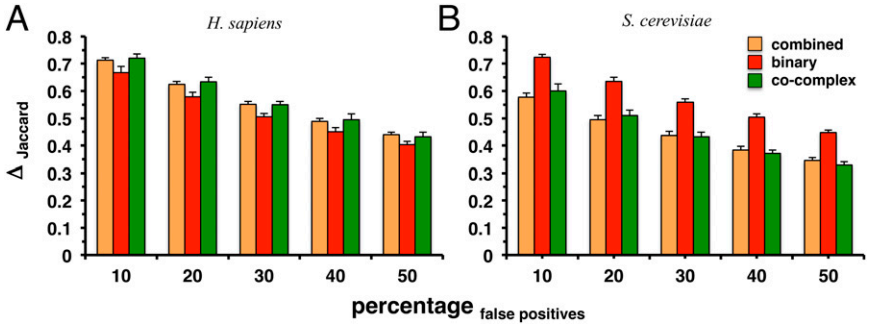

Fig. S3. Robustness of MDSets. We compared the MDSet of the actual and randomized networks by a Jaccard index, defined as $\Delta_{ij} = N_{ij}/(N_i + N_j - N_{ij})$, where N_{ij} was the number of common MDSet proteins in networks i and j, and N_i was the number of MDSet proteins in network i. To represent false-positive interactions we randomly deleted the corresponding percentage of interactions in the underlying combined, binary, and cocomplex interaction sets in (*A*) human and (*B*) yeast.

A

	H. sapiens		
	combined	binary	co-complex
cancer-related	329 (232) 9.9×10^{-43}	306 (237) 2.8×10^{-31}	232 (164) 1.2×10^{-25}
virus-targeted	280 (180) 8.5×10^{-57}	253 (186) 1.1×10^{-40}	199 (132) 3.6×10^{-24}

	S. cerevisiae		
	combined	binary	co-complex
essential	303 (190) 1.6×10^{-51}	216 (228) 5.9×10^{-6}	280 (167) 4.7 $\times 10^{-47}$

B

	H. sapiens		
	combined	binary	co-complex
cancer-related	306 (232) 2.1×10^{-32}	288 (237) 2.1×10^{-24}	219 (164) 6.4×10^{-20}
virus-targeted	272 (180) 9.2×10^{-52}	261 (186) 4.2×10^{-45}	201 (132) 3.0×10^{-25}

	S. cerevisiae		
	combined	binary	co-complex
essential	244 (190) 5.0×10^{-22}	224 (228) 1.3×10^{-7}	212 (167) 9.6×10^{-14}

Fig. S4. Enrichment of hubs among disease and essential genes. (*A*) We calculated how often cancer-related genes, proteins that are targeted by human viruses, and essential genes appeared in sets of most connected proteins that equaled the size of the corresponding MDSet in the given human and yeast interaction sets. Using Fisher's exact test we determined the significance of such functional proteins in these sets of most connected proteins. In parentheses we show the corresponding numbers in the matching MDSets. In *B* we present the corresponding numbers using sets of proteins with highest betweenness centrality that matched the corresponding MDSets in the underlying interaction networks.

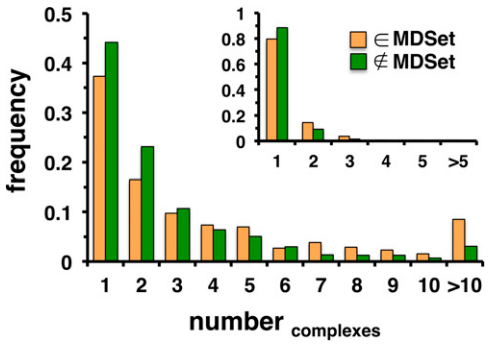

Fig. S5. Involvement of MDSet proteins in protein complexes. Counting the number of complexes proteins are involved in, we observed that human MDSet proteins in the combined human interaction network appeared in more complexes than non-MDSet proteins ($P = 2.6 \times 10^{-8}$, Wilcoxon test). (*Inset*) We observed a similar result using (non)MDSet proteins in the combined yeast interaction network ($P = 4.1 \times 10^{-5}$).

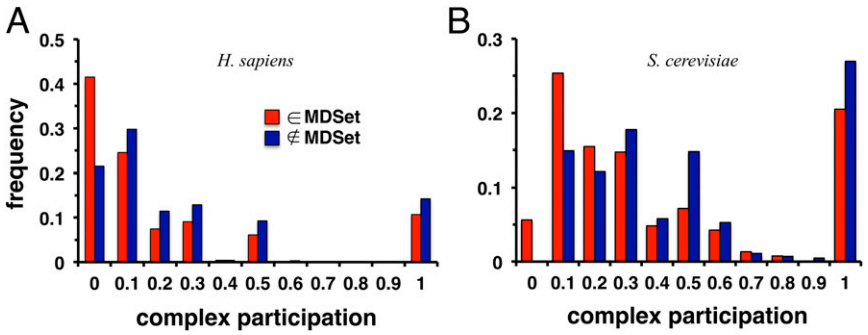

Fig. S6. Complex participation of MDSet proteins. (*A*) Focusing on the combined human network, we found that human MDSet proteins have significantly lower complex participation values than non-MDSet proteins ($P = 2.4 \times 10^{-38}$, Wilcoxon test). In (*B*) we found a similar result for (non)MDSet proteins in the combined yeast interactions set ($P = 3.3 \times 10^{-17}$).

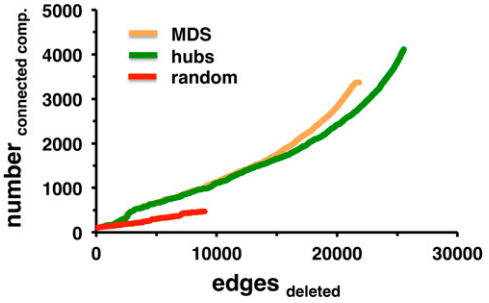

Fig. S7. Robustness of underlying human network. We sorted all MDSet proteins in the combined human interaction network according to their degree. As an equivalent set of equal size we collected and sorted the highest connected hub proteins. Furthermore, we randomly sampled a set of proteins of the same size. Starting with the most connected protein, we gradually deleted proteins and calculated the number of connected components in the altered network.

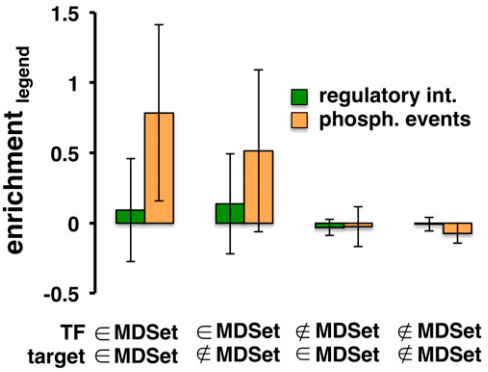

Fig. S8. Enrichment of MDSet proteins among yeast transcription factors, kinases, and their targets. We calculated the enrichment of transcription regulatory interactions and phosphorylation events between (non)MDSet proteins in the combined yeast interaction network. We observed that both types of interactions preferably appeared between MDSet proteins, whereas we found the opposite for non-MDS proteins. Specifically, we obtained best results when the corresponding transcription factor or kinase was involved in the MDSet.

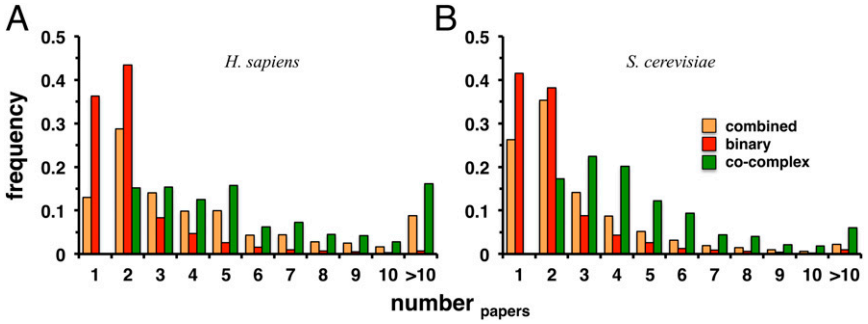

Fig. S9. Quality analysis of protein interaction sets. We counted the number of papers that reported the presence of an interaction. In (*A*) human and (*B*) yeast we observed that the majority of interactions in the binary interactions sets were confirmed in more than one paper. Cocomplex interactions were only accounted for when they appeared in at least two publications.

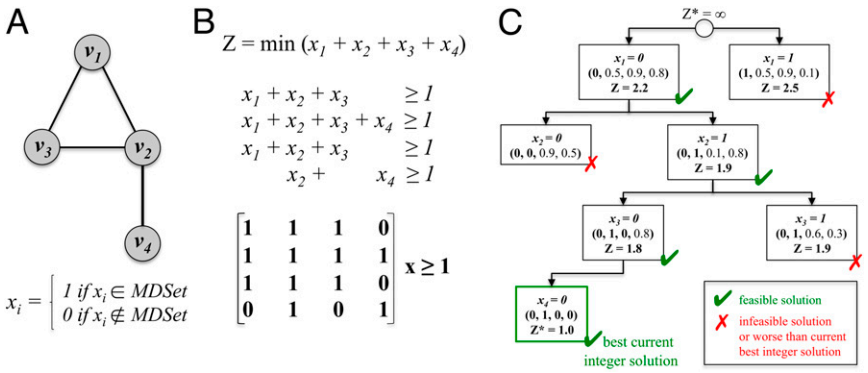

Fig. S10. Concept and solution of binary integer programming problems. (*A*) Toy network to illustrate the determination of an MDSet. (*B*) Formulation of the binary integer programing problem to determine the MDSet of the toy network. (*C*) Search tree as generated by a branch-and-bound algorithm to solve the underlying binary integer programming problem to find an MDSet.