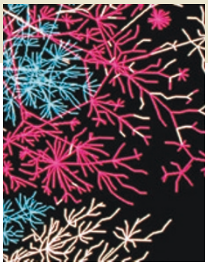


Effect of sampling on topology predictions of protein-protein interaction networks

Jing-Dong J Han[1–3], Denis Dupuy[1,3], Nicolas Bertin[1], Michael E Cusick[1] & Marc Vidal[1]

Currently available protein-protein interaction (PPI) network or 'interactome' maps, obtained with the yeast two-hybrid (Y2H) assay or by co-affinity purification followed by mass spectrometry (co-AP/MS), only cover a fraction of the complete PPI networks. These partial networks display scale-free topologies–most proteins participate in only a few interactions whereas a few proteins have many interaction partners. Here we analyze whether the scale-free topologies of the partial networks obtained from Y2H assays can be used to accurately infer the topology of complete interactomes. We generated four theoretical interaction networks of different topologies (random, exponential, power law, truncated normal). Partial sampling of these networks resulted in sub-networks with topological characteristics that were virtually indistinguishable from those of currently available Y2H-derived partial interactome maps. We conclude that given the current limited coverage levels, the observed scale-free topology of existing interactome maps cannot be confidently extrapolated to complete interactomes.

Determining the topology of a network, the configuration of its nodes and the connecting edges, is relevant for assessing network stability, dynamics and function, and ultimately for being able to design and reengineer networks of interest[1] (see **Box 1**). Only recently has it become possible to discern the topology of large, complex networks[1,2]. Typically, such networks and their topologies are determined using a variety of sampling methods. These partial networks are then used to infer the topology of the whole network[3,4].

Existing large-scale protein-protein interaction (PPI) or interactome network maps are considered scale free[5–10]. To date two methods have been used to generate such maps: yeast two-hybrid (Y2H)[5–9] and co-affinity purification followed by mass spectrometry (co-AP/MS)[11,12]. Despite the wealth of information collected in these maps it is important to remember that they are partial maps covering only a small fraction of the total *in vivo* interactome[13–18]. The total number of interactions in the yeast *Saccharomyces cerevisiae* interactome has been estimated to be 15,000–30,000 (refs. 13,17). Maps derived from the two published co-AP/MS data sets[11,12] contain putative interactions, predicted on the basis of co-membership in a protein complex. Thus, a significant fraction of the PPIs reported in each of these maps may correspond to indirect interactions that would lead to a significant overestimation of their actual coverage. The yeast interactome maps generated by analyzing direct binary interaction assays with the Y2H assay independently cover a mere 3–9% of the complete interactome (948 and 806 defined in the Uetz and Ito core Y2H maps respectively[5,6]) (**Table 1**). The *Caenorhabditis elegans* and *Drosophila melanogaster* Y2H interactome maps show similar limited coverage[8,9]. This low coverage can explain the limited overlap observed between large-scale yeast Y2H data sets[5,6,14,18], between large-scale co-AP/MS data sets[11,12,18] and between *D. melanogaster* Y2H data sets[8,19] (see **Box 2**). To extrapolate the topology of complete interactomes from such incomplete maps requires the assumption that the limited sampling does not affect the overall topological analyses[20]. Recent reports have already noted discrepancies in matching existing interactome networks to a scale-free topology[21,22].

Here, we analyze whether extrapolation of network topologies from partial network data to the whole network can be done accurately and with high confidence. Our approach consisted of generating theoretical, topology models, sampling them without introducing erroneous interactions (no false positives), analyzing the resulting network topologies and comparing them to experimental data. This approach is in contrast to that of previous studies, where experimental PPI data sets are directly matched to a topology model[10,22–24]. We show with these *in silico* simulations that limited sampling alone can give rise to apparent scale-free topologies, irrespective of the original network topology, and thus complete network topologies cannot be extrapolated directly from sub-network data.

Sampling complete theoretical interactomes

The two commonly used PPI mapping approaches have fundamental technical differences that preclude the use of a single method to simulate them both. On the one hand, co-AP/MS detects co-membership in a protein complex. A co-AP/MS complex can thus contain proteins that are second-degree and third-degree interactors of the protein used as bait. In co-AP/MS-derived interactome maps the edges between the proteins are therefore predicted interactions, generated by either the 'spoke' model (the bait is predicted to interact directly with all members of a complex) or the 'matrix' model (all members of the complex are presumed to interact directly with all other members of the complex)[11,12,15]. On the other hand, Y2H detects only direct binary interactions. Our sampling simulation was designed to model such direct binary PPI assays only (see **Box 3**).

[1]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, 44 Binney Street, Boston, Massachusetts 02115, USA. [2]Present address: Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Road, Beijing, 100101, China. [3]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu).

Published online 7 July 2005; doi:10.1038/nbt1116

Box 1 Why topology matters

The topology of a network refers to the relative connectivity of its nodes. Different topologies affect different specific network properties.

Scale-free networks are resistant to random failure but vulnerable to targeted attack, specifically against the most connected proteins (hubs)[36]. This property has been held to account for the robustness of biological networks to perturbations like mutation and environmental stress[1,22,37]. A positive correlation between essentiality and connectivity has been demonstrated, linking topological centrality to functional essentiality[22]. Identification of high-degree proteins would then represent one strategy for therapeutic mediation of signaling pathways that go awry in cancer[38]. Investigating high-degree proteins as drug targets might prove a valuable approach[1,39]. Such a strategy would have lesser impact if the true topology were exponential, and would be inoperable if the true topology were random.

Current interactome maps are far from complete[13–18]. Since high-coverage interactome mapping is costly, a strategy for interactome mapping optimization is desirable. One proposed strategy assumes that the complete interactome is scale free. The strategy focuses on an iterative process where the identified hubs are used as baits in the subsequent mapping step[40]. If the topology of the complete interactome were not truly scale free then this strategy would be less cost effective.

How the interactome evolved into its present form is naturally a question of great interest. Numerous hypotheses on the mechanism underlying the evolution of interactomes are based on their topology[41–43]. Different models produce different topologies. To decide which model comes closest to reality calls for accurately defining the true topology of the full interactome.

Whether the interactome is organized into modules, clusters of interconnected proteins that have related or identical biological function, is another question of compelling interest[10,44–46]. Various methods for dividing the interactome into modules have been presented[10,47–51]. To evaluate the relative performance of these approaches it helps to have an understanding of the true topology of the full interactome.

We applied our simulation procedure to four different starting network topologies: (i) random networks as defined by Erdös & Rényi (ER)[25] for which the distribution of the number of nodes of a given degree (number of interactions) follows a Poisson distribution

$$f(k) = e^{-m}\left(\frac{m^k}{k!}\right),$$

where m is the average number of occurrencer per k. When $m >> 1$ it can be approximated by a normal distribution

$$(f(k) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{-(k-\mu)^2}{2\sigma^2}\right)}, \text{ with } \sigma < \mu)$$

where σ is the standard deviation and μ the mean of the degree distribution); (ii) exponential networks (EX) whose degree distribution follows an exponential function ($f(k)=e^{-\gamma k}$); (iii) scale-free networks whose degree distribution follows a power law (PL) ($f(k)=k^{-\gamma}$)[26]; (iv) networks that have a normal distribution with a standard deviation greater than the mean that lead to a dramatic left truncation of the distribution (truncated normal or TN)

$$(f(k) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{-(k-\mu)^2}{2\sigma^2}\right)}, \text{ with } \sigma > \mu)$$

(**Supplementary Fig. 1** and **Supplementary Note** online). For each of these topologies we generated networks containing 6,000 nodes (the predicted size of the yeast proteome[27]) with an average degree (<*k*>) of 5, 10 and 20.

Table 1 Topological properties of interactome maps

Data set	Ito *et al.* (yeast)	Uetz *et al.* (yeast)	Ito-Uetz combined	Li *et al.* (worm)	Giot *et al.* (fly)	Minimum value	Maximum value
Total number of nodes	797	1,005	1,417	1,415	4,651	797	4,651
Nodes in main component	417 (52%)	473 (47%)	970 (68%)	1,260 (89%)	3,039 (65%)	47%	89%
Total number of interactions	806	948	1,520	2,135	4,787	806	4,787
Interactions in main component	544	558	1,229	2,038	3,715	544	3,715
R-square	0.843	0.954	0.899	0.885	0.91	0.843	0.954
γ	–1.82	–2.42	–1.91	–1.59	–2.75	–2.75	–1.59
<*k*>	1.96	1.84	2.15	2.98	2.04	1.84	2.98
Average clustering coefficient	0.2	0.11	0.09	0.09	0.06	0.06	0.2
Number of network components	143	177	160	70	591	70	591
Average component size	5.6	5.7	8.9	20.2	7.9	5.6	20.2
Characteristic path length	6.14	7.48	6.55	4.91	9.43	4.91	9.43
Number of baits	455	512	827	502	2,820	455	2,820

The linear regression R-square measures the linearity between log(n(*k*)) and log(*k*) i.e. the fit to a power-law distribution. γ is the exponent of the power law distribution formula that best fits the observed distribution. <*k*> is the average number of interactions per protein observed in the network. For the Ito, Li and Giot data sets only the high confidence interactions were considered (core).

Box 2 Coverage versus accuracy

False positives (identified interactions that do not occur physiologically) in PPI maps are of two classes[9,33–35]. On the one hand, technical false positives arise from limitations of the experimental procedures used—the more stringent implementations of Y2H recently developed allow elimination of most of these[9,33–35]. On the other hand, biological false positives are PPI that occur in the experimental procedure but do not occur *in vivo*, because the two proteins are not expressed at the same time, in the same sub-cellular compartment, or in the same tissue. We chose not to introduce spurious interactions into our sampling simulation. First, we wanted to study the impact of sampling alone on the topological features of the map. Second, as false-positive interactions in the PPI maps are not recognizable (otherwise they would have been eliminated) any attempt to simulate them *a priori* would have to rely on speculative assumptions. Lastly, this approach permits comparison between samples containing only true positives and the experimental data, thereby providing insight into the impact of false positives on the observed topology.

The minimal overlap between independently generated Y2H data sets has led to the supposition that these data sets are skewed by false-positive interactions[16,17]. However, as each map covers a mere 3–9% of the total interactome, it can be argued that this poor overlap is to be expected. We systematically compared all sampled networks of a size comparable to the Uetz *et al.* yeast Y2H data set[5] (black dots in **Fig. 2**) to all samples of a size comparable to the Ito-Core data set[6] (black dots in **Supplementary Fig. 2a** online). In 23,903 such comparisons, the average fraction of interactions that were common to each pair of sampled maps was 2.1% (±0.94%) with a maximum of 7.8%. This result indicates that it is possible to observe perfectly accurate samples (without false positives) that have very limited overlap solely because of the low coverage of these maps. Therefore, the low overlap between Y2H maps does not allow firm conclusions regarding their accuracy, and the experimental Y2H data sets may be of better quality than is ordinarily believed.

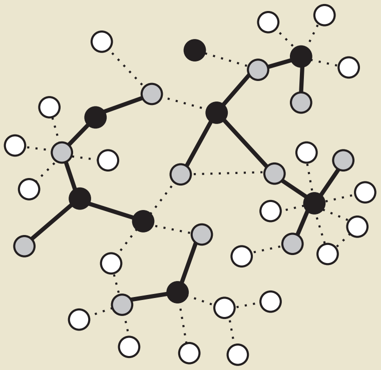
The dramatic impact of sampling on the observed topology is illustrated with an ER network with <k> = 10: the more limited the sampling, the more the peak of the degree distribution shifts toward a lower degree (**Fig. 1**). Ultimately only a truncated exponential decay tail is evident, whose shape resembles a scale-free degree distribution with an R-square value approaching 1 (**Fig. 1c**).

Low coverage sampling of all four of these network types generated maps whose degree distribution could be approximated by a power-law function as reflected by the R-square evaluations of linearity (**Fig. 2**, orange-red region of each panel). Marginal coverage of all four model topologies can thus give rise to sampled networks whose distribution fits a power law function as well as the available experimental interactome networks do (**Table 1**). We repeated this experiment with networks of 20,000 nodes, simulating the size of the worm *C. elegans*[28] or the fly *D. melanogaster*[29] proteomes, and obtained similar results (**Supplementary Fig. 2** online, **Supplementary Table 1** online).

Comparison of existing interactome maps with *in silico* maps

Although degree distribution is one of the most widely studied network parameters, it does not capture all aspects of the topology of a network[2,24,26]. To further test if the sampled networks obtained in our simulations resemble current interactome maps, we assessed other topological parameters that follow a common trend across all available experimental data sets (**Table 1**). These parameters include: (i) the

Box 3 Sampling methodology



To simulate the sampling that occurs in binary PPI assays mapping experiments, such as Y2H, we derived two metrics, bait coverage and edge coverage. Bait coverage simulates the percentage of proteins in the entire proteome experimentally tested as baits in large-scale mapping experiments. Edge coverage simulates the limited percentage of interactions per bait recovered in large-scale mapping experiments, arising from both technical and experimental limitations[9,33–35]. Preys are also randomly selected from the entire proteome. Bait can be preys and preys can also be baits, just as in actual Y2H screens. For each theoretical topology model tested we scanned through the full range of bait and edge coverage each from 0 to 100%. For each combination of bait and edge coverage, we generated a sampled network by first randomly picking nodes as baits (black nodes in diagram) at the designated bait coverage, then randomly selecting as preys (grey nodes in diagram) a fraction of their interactors, according to the designated edge coverage. Both the original theoretical network (dotted lines in diagram) of various topologies and the sampled networks (solid lines in diagram) are undirected graphs, that is, edges A-B and B-A are regarded as one single edge and homodimeric interactions are allowed and regarded as one single edge. We then examined how closely the degree distributions of the sampled networks so obtained match a scale-free distribution. In a scale-free network the degree distribution follows a power law distribution[3]. The linear regression R-square function was used to assess linearity between log(n(k)) and log(k). R-square ranges from 0 to 1, with 1 representing perfect linearity, that is, a perfect power-law distribution.

Given estimations of average k (<k>) in full yeast interactomes[13,17], we chose to model the complete interactomes with <k> of 5, 10 and 20. Networks matching a predefined degree distribution formula for ER, PL, TN or EX distribution were generated by an edge allocation algorithm. Briefly, we first allocate the number of links each node in the network makes according to the degree distribution formula, then we randomly pick a pair of nodes to make an edge, then decrease the degree for both nodes by one until the desired number of edges has been assigned to nodes.

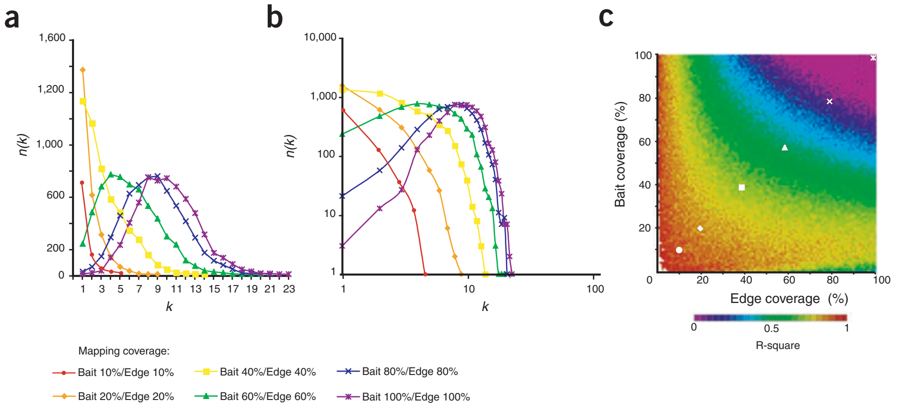


Figure 1 Sampling of an Erdös-Rényi random network. (**a**) Degree distributions of sampled networks, starting from a random network of <*k*> = 10. The bait and edge coverage corresponding to each curve is marked by a white dot in **c**. (**b**) Degree distributions of the sampled networks on log-log scale. The bait and edge coverage corresponding to each curve is marked by a white dot in **c**. (**c**) The linear regression R-square function is used to measure the linearity between log(n(*k*)) and log(*k*). The colors corresponding to particular R-square values are plotted against discrete bait and edge coverage of network sampling. The red-orange end of the color scale indicates strong linearity, while the purple-blue end of the color scale indicates poor linearity.

value of the log(n(*k*)) and log(*k*) linear regression R-square (≥ 0.843); (ii) the value of the exponent γ of the power law distribution formula $n(k) \approx k^{-\gamma}$ (between 1.59 and 2.75); (iii) the fraction of nodes in the main component[24,30] (between 0.471 and 0.89); and (iv) the average degree *k* of the networks (between 1.84 and 2.98). Sampled networks for which all four parameters are within the described ranges are shaded in each panel in **Figure 2**. For all four network topologies tested there were regions of coverage that satisfy all the topological constraints (shaded region in **Figure 2**). Although the ER graphs have significantly smaller areas satisfying the constraints than the other topologies tested, this does not necessarily exclude a random distribution as the underlying topology of the PPI networks.

To investigate how closely our simulation could approach the experimental observations, we decided to examine samples that have a size similar to that of the available data sets. We present here the examination of simulated samples that correspond in size to the Uetz *et al.* yeast Y2H data set[5] (1005 proteins and 948 interactions). This data set has similar accuracy[16] to but is slightly larger than Ito-Core[6], the other available large-scale yeast Y2H data set. Results for the other currently available binary PPI data sets[5–9] (**Table 1**) are reported (**Supplementary Figs. 2** and **3** online and **Supplementary Table 1** online). Networks of similar size are defined as those sampled networks whose number of nodes and edges are within 10% of those of the corresponding PPI map, and are marked by black dots in **Figure 2**. We obtained sampled networks of similar size to four out of the five Y2H data sets examined. For several of the Y2H maps these sampled networks also fit the four consensus topological features described above (**Fig. 2** and **Supplementary Fig. 2** online).

For each of the 12 (three <*k*> values and four distributions) starting *in silico* networks (6,000 nodes), we plotted the degree distribution of the sampled networks of comparable size, overlaying them onto the degree distribution of the Uetz *et al.* Y2H map (**Fig. 3**). Many of the selected networks sampled from PL, TN and EX topologies display a degree distribution that is consistent with that of the Uetz *et al.* map. A few sampled networks derived from the ER5 and ER10 networks also fit this distribution, but they generally display a sharp drop-off at the higher degree end. Similar fitting trends were observed when the comparison was carried out with other Y2H data sets (**Supplementary Table 1** online and **Supplementary Fig. 3** online). If the number of

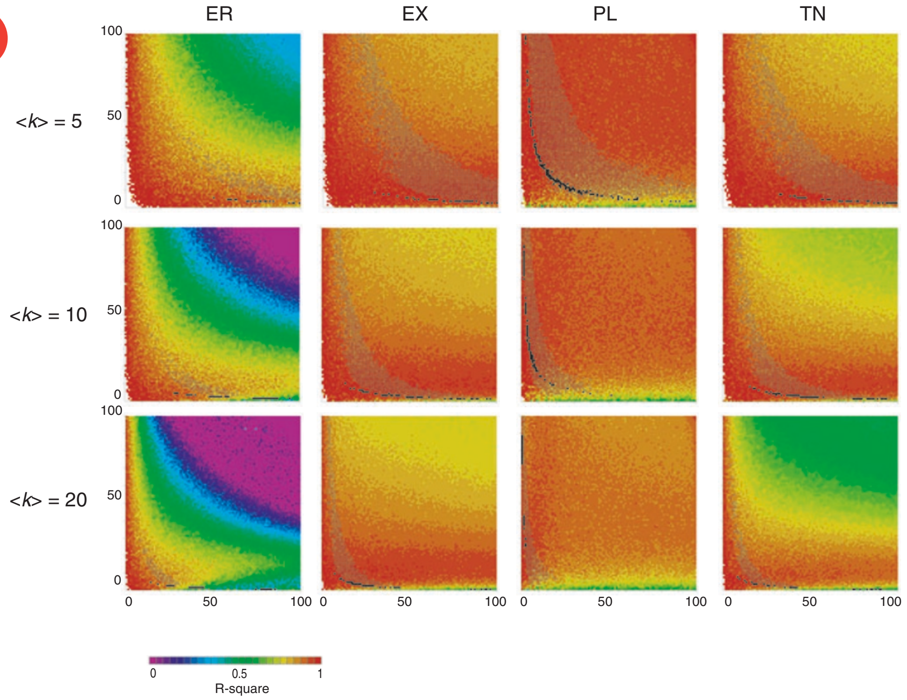
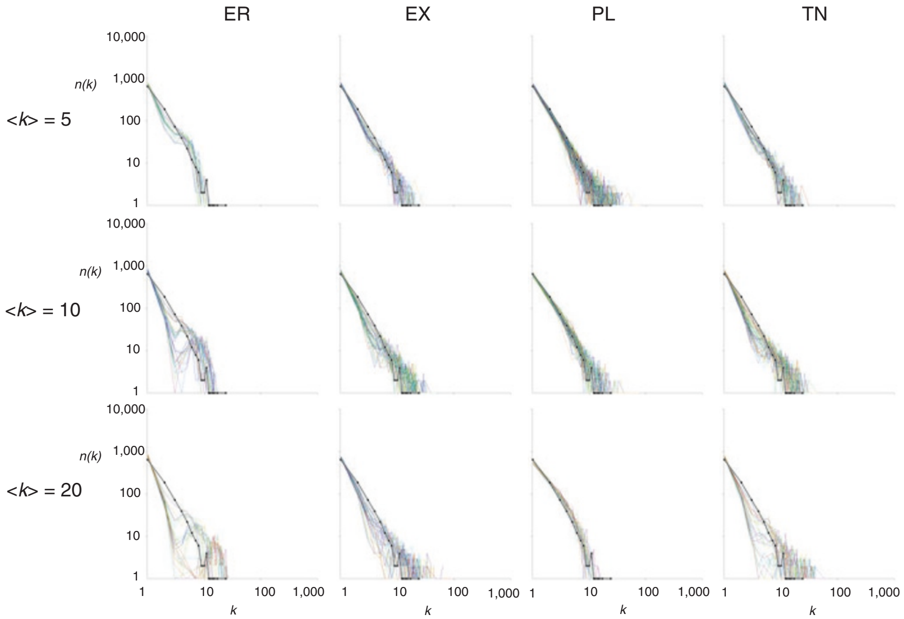


Figure 2 Sampled networks derived from starting networks of various topologies. The simulations are based on the size of the yeast proteome (6,000 nodes). The starting networks are Erdös-Rényi random networks (ER), exponential networks (EX), scale-free networks that follow a strict power law distribution (PL), and networks having a truncated normal distribution (TN). Each network has average degree <*k*> values of 5, 10 and 20. The networks are sampled at discrete bait and edge coverage. For each combination of bait coverage (y coordinate) and edge coverage (x coordinate) the color represents the value of the linear regression R-square function between log(n(*k*)) and log(*k*) of the resulting sampled network. The shaded regions in each panel indicate the regions of bait and edge coverage that fit all four topological constraints described in the text. (**Table 1**). Black dots indicate that the sample obtained at the corresponding bait and edge coverage was of a size similar to the Uetz *et al.* PPI data set[5]. The axes are as in **Figure 1c**.

Figure 3 Degree distribution of sampled networks. The Uetz *et al.* PPI[5] data set (black line) is compared to sampled networks (colored lines) of similar size, indicated by the black dots in **Figure 2**. The node degree *k* is represented on the x-axis, and the number of nodes with a particular *k* is represented on the y-axis.



sampled networks of comparable size to experimental Y2H maps is taken to indicate the likelihood of a particular underlying topology, then the PL model seems more likely than the other tested models. However, it is not possible to definitely exclude any of the other tested topologies based solely on the existing interaction data.

Conclusions

A recent report used mathematical modeling to demonstrate a distorting impact of sampling on the apparent topology of scale-free networks[20]. Our simulation raises the possibility that the apparent scale-free topology of the experimental Y2H interactome maps[5–9] may not represent the true topology of the full interactome. Current bait and edge coverage in these data sets is so limited that none of the topologies tested herein can be definitively ruled out. The difficulty presented by marginal sampling is not limited to Y2H but would arise if any other method for determination of binary interactions, such as protein microarrays[31], binary co-affinity purification[9], or phage display[32], were applied in large-scale experiments.

We have shown that limited sampling alone can lead to misleading degree distribution, apart from any influence of the false positives. Limited sampling can also explain the lack of overlap between independent maps, which is often attributed to 'noisy' data sets (see **Box 2**). Many technical false positives are auto-activators or sticky proteins represented by nodes of artificially high degree[33–35], which might tilt the apparent topology even more towards scale free than that observed here by sampling alone.

Our study does not imply that scale-free topology is always an artifact of sampling. The scale free Internet and World Wide Web networks are necessarily sampled because of their vast size[3,4,26]. Even so, they are likely truly scale free, because although node coverage is very low, edge coverage by the sampling methods used is close to 100 percent (bottom right corner in the panels of **Fig. 2**).

Lastly, it is little appreciated just how low the coverage of the complete interactome is in existing interactome maps[18]. Our results show that low coverage makes a determination of the true topology of the interactome difficult, indicating a dire need to increase coverage through further experimentation, as well as through development of improved PPI mapping technology[9,31–35].

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS
We thank Mike Boxem, Fritz Roth and Debra Goldberg for extremely useful discussions and careful reading of the manuscript. This work was supported by grants from NIGMS, NHGRI, and NCI to M.V.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/

1. Barabási, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
2. Strogatz, S.H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
3. Barabási, A.L., Albert, R. & Jeong, H. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A (Amsterdam)* **281**, 69–77 (2000).
4. Yook, S.H., Jeong, H. & Barabasi, A.L. Modeling the Internet's large-scale topology. *Proc. Natl. Acad. Sci. USA* **99**, 13382–13386 (2002).
5. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae. Nature* **403**, 623–627 (2000).
6. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
7. Reboul, J. *et al. C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**, 35–41 (2003).
8. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster. Science* **302**, 1727–1736 (2003).
9. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans. Science* **303**, 540–543 (2004).
10. Han, J.D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
11. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
12. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
13. Walhout, A.J., Boulton, S.J. & Vidal, M. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* **17**, 88–94 (2000).
14. Edwards, A.M. *et al.* Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **18**, 529–536 (2002).
15. Bader, G.D. & Hogue, C.W. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.* **20**, 991–997 (2002).
16. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
17. Grigoriev, A. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.* **31**, 4157–4161 (2003).
18. Ito, T. *et al.* Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell. Proteomics* **1**, 561–566 (2002).
19. Formstecher, E. *et al.* Protein interaction mapping: A *Drosophila* case study. *Genome Res.* **15**, 376–384 (2005).
20. Stumpf, M.P., Wiuf, C. & May, R.M. Subnets of scale-free networks are not scale-free:

Sampling properties of networks. *Proc. Natl. Acad. Sci. USA* (2005).

21. Przulj, N., Corneil, D.G. & Jurisica, I. Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508–3515 (2004).
22. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
23. Thomas, A., Cannings, R., Monk, N.A. & Cannings, C. On the structure of protein-protein interaction networks. *Biochem. Soc. Trans.* **31**, 1491–1496 (2003).
24. Yook, S.H., Oltvai, Z.N. & Barabasi, A.L. Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928–942 (2004).
25. Erdös, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–60 (1960).
26. Barabási, A.L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
27. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–547 (1996).
28. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
29. Adams, M.D. *et al.* The genome sequence of *Drosophila melanogaster. Science* **287**, 2185–2195 (2000).
30. Schwikowski, B., Uetz, P. & Fields, S. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257–1261 (2000).
31. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
32. Tong, A.H. *et al.* A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–324 (2002).
33. Walhout, A.J. & Vidal, M. A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Res.* **9**, 1128–1134 (1999).
34. Legrain, P., Wojcik, J. & Gauthier, J.M. Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet.* **17**, 346–352 (2001).
35. Vidalain, P.O., Boxem, M., Ge, H., Li, S. & Vidal, M. Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* **32**, 363–370 (2004).
36. Albert, R., Jeong, H. & Barabasi, A.L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
37. Wagner, A. Robustness against mutations in genetic networks of yeast. *Nat. Genet.* **24**, 355–361 (2000).
38. Vogelstein, B., Lane, D. & Levine, A.J. Surfing the p53 network. *Nature* **408**, 307–310 (2000).
39. Apic, G., Ignjatovic, T., Boyer, S. & Russell, R.B. Illuminating drug discovery with biological pathways. *FEBS Lett.* **579**, 1872–1877 (2005).
40. Lappe, M. & Holm, L. Unraveling protein interaction networks with near-optimal efficiency. *Nat. Biotechnol.* **22**, 98–103 (2004).
41. Eisenberg, E. & Levanon, E.Y. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **91**, 138701 (2003).
42. Qin, H., Lu, H.H., Wu, W.B. & Li, W.H. Evolution of the yeast protein interaction network. *Proc. Natl. Acad. Sci. USA* **100**, 12820–12824 (2003).
43. Pereira-Leal, J.B., Audit, B., Peregrin-Alvarez, J.M. & Ouzounis, C.A. An exponential core in the heart of the yeast protein interaction network. *Mol. Biol. Evol.* **22**, 421–425 (2004).
44. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
45. Poyatos, J.F. & Hurst, L.D. How biologically relevant are interaction-based modules in protein networks? *Genome Biol.* **5**, R93 (2004).
46. Bork, P. *et al.* Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* **14**, 292–299 (2004).
47. Dunn, R., Dudbridge, F. & Sanderson, C.M. The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics* **6**, 39 (2005).
48. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli. Nat. Genet.* **31**, 64–68 (2002).
49. Rives, A.W. & Galitski, T. Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA* **100**, 1128–1133 (2003).
50. Bader, G.D. & Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
51. Pereira-Leal, J.B., Enright, A.J. & Ouzounis, C.A. Detection of functional modules from protein interaction networks. *Proteins* **54**, 49–57 (2004).