

Protein-protein interaction networks and biology—what's the connection?

Luke Hakes, John W Pinney, David L Robertson & Simon C Lovell

Analysis of protein-protein interaction networks is an increasingly popular means to infer biological insight, but is close enough attention being paid to data handling protocols and the degree of bias in the data?

The availability of large-scale protein-protein interaction data has led to the recent popularity of the study of protein interaction networks. Just as the enormous amount of available sequence data has made it possible to obtain an overview of the genome, it is hoped that interaction data will allow an analogous view of the interactome. However, we urge caution. Although there are plenty of data, knowledge of the interactome remains incomplete. More importantly, the available data contain biases as a consequence of differences in the origins and handling of the different datasets. These biases can alter the underlying structure of the network in unpredictable ways. Paradoxically, some attempts to increase data quality make these biases more severe. As a result, inferring meaningful biological conclusions from network topology remains problematic.

The rise of interactomics

Technological advances in proteomics have resulted in unprecedented quantities of protein-protein interaction (PPI) data. This is particularly true for the yeast *Saccharomyces cerevisiae*^{1–4}, where the most comprehensive studies have been performed. In parallel with this huge increase in data, developments in network theory—for example, the ubiquitous identification of scale-free networks⁵—have led to the application of network-based analyses to PPIs. Our current view of the set of yeast PPIs is as a large, interlinked set of interactions between proteins represented as nodes in a graph^{6,7}. This has stimulated research into the

properties of the PPI network, with a view to understanding the biological properties of the systems that underlie them.

The topology of the network potentially has a wide range of biological implications. For example, it has been suggested that the structure of the PPI network is related to whether or not a given protein is essential^{6,8,9}, that hub-centric modularity arises naturally from network topology¹⁰ and that the scale-free structure of observed networks implies robustness with respect to random component failure^{11,12}. Analysis of network structure has been used to propose evolutionary mechanisms of how cellular complexity arose^{13–16} and has also led to controversy as to whether network modularity is dynamically organized^{8,17–19}.

The prospect of inferring biological conclusions from network structure is part of what makes these interaction data so interesting and important. However, we suggest that caution is warranted. Although it is widely known that there are problems with data quality in terms of missing interactions and false positives²⁰, issues related to sampling bias are less well understood. Here, we argue that the effects of data handling protocols and the degree of bias in the data are at least as problematic as data quality issues, perhaps to the point that no large-scale parameters can be trusted.

Sampling and biases

Current estimates of the proportion of known protein-protein interactions suggest that in yeast, ~50% of interactions have been identified²¹. Estimates of known PPIs for humans are much lower, at ~10% of the complete set²¹. Current PPI networks are, therefore, a sample of the complete network. Ideally, this sample would be unbiased. However, even with random sampling, incompleteness has an enor-

mous effect on overall network topology^{22–24}. In fact, the set of known interactions is even less representative of the whole network than these studies would suggest, because the subset of interactions that have been identified is by no means random. Biases in sampling lead to even more drastic differences between the complete network and the subsample that we observe. Even those data derived from high-throughput studies are not an unbiased sample of the complete network; rather, they are biased toward proteins from particular cellular environments, toward more ancient, conserved proteins and toward highly expressed proteins²⁰.

In an attempt to circumvent the problem of inaccurate data²⁰, several authors have tried to produce both more complete and more reliable datasets. The most extensive collation of literature curated data is the 'LC' dataset of Reguly *et al.*²⁵. This dataset largely consists of proteins studied either individually or in small-scale studies. It therefore differs in two ways from datasets derived from high-throughput studies. First, these proteins have been selected for careful study by biologists. This means that they are more likely to have been studied before, to be disease related or, in yeast, to be essential proteins. Essential proteins also tend to be more highly connected than nonessential proteins⁶, although it is unclear whether this is an intrinsic property (for example, a consequence of the correlation between connectivity, essentiality and evolutionary age of a protein²⁶) or a function of their having been more thoroughly studied, or a combination of the two. Second, there are systematic methodological differences between interactions identified through small- and large-scale studies. An obvious difference is the increased reliability, and thus lower rate of false positives, in small-scale studies resulting from the

Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Rd., Manchester M13 9PT, UK.
e-mail: simon.lovell@manchester.ac.uk

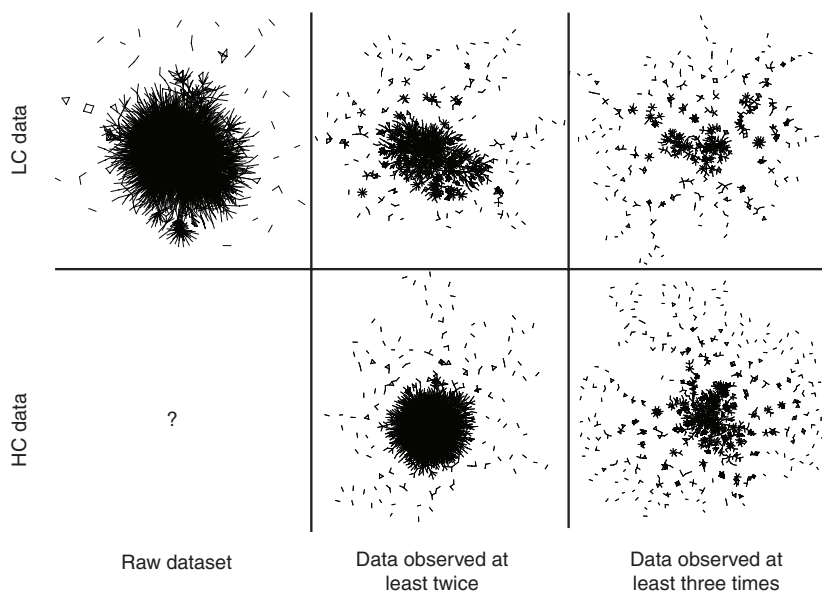


Figure 1 Visualization of the yeast protein-protein interaction network showing the effect of multiple validation on networks derived from both the LC (Reguly *et al.*²⁵) and the HC (Batada *et al.*⁷) datasets. The HC dataset contains only those interactions observed at least twice; the data it was derived from are not available. Networks were visualized with LGL³⁴.

fact that these types of experiments tend to have more careful controls.

So-called study bias can have marked effects on biological conclusions drawn from interaction data²⁷. Pereira-Leal *et al.*²⁸ found that there is a difference in connectivity between yeast essential proteins and the network taken as a whole, and attributed this as being a property of the ancestral network. This difference in connectivity can be identified by analysis of the ‘degree correlation’. The degree of a node in a network is the number of connections it has; in the case of a PPI network, this represents the number of other proteins to which a given protein can bind. A negative correlation between a protein and its neighbors in the network was observed for the full yeast dataset, indicating that highly connected proteins tend to be isolated from each other¹⁰. This is unusual as, in most networks, highly connected nodes tend to interact with other highly connected nodes²⁹. Any correlation in degree can be identified by plotting the connectivity of a protein against the average connectivity of its neighbors in the network. Thus, it was suggested that the yeast interaction network as a whole shows a negative correlation in degree, indicating hub-centric modularity¹⁰, whereas the set of essential proteins shows no such correlation²⁸.

Later work demonstrated that data selection is the prime determinant of degree correlation²⁷. Indeed, it is possible to distinguish data selection effects from genuine biological

effects by controlling for study size. When this was done, it was found that the difference between essential and nonessential proteins was no longer present²⁷. However, by careful choice of the source of data about essential proteins (for example, by including only those data derived from yeast two-hybrid assays or only those found in large protein complexes), it is possible to detect a negative correlation between the connectivity of neighboring nodes²⁷. Additionally, a change from the ‘spoke’ to the ‘matrix’ model for decomposing interactions from complexes into individual sets of interactions changes the degree correlation from negative to positive. It seems, then, that data handling is of primary importance, as it is this—not biological differences—that most influences degree correlation.

Combination of datasets and multiple validation

As well as being incomplete, the available data are notoriously unreliable. Analysis of the initial high-throughput datasets^{1,2} has shown that there is little overlap between them²⁰. This is due to a combination of both incompleteness and the high number of false positives. More recent datasets^{3,4} have improved in both quality and coverage. However, missed interactions and incorrectly identified interactions both remain problems.

There have been several attempts to increase the degree of confidence that may be ascribed

to PPIs. Many of these approaches use multiple validation—the rationale being that interactions observed multiple times are more likely to be true than those that have only been observed once. Assigning different weights to interactions derived from different types of experiments has also been proposed (see Suthram *et al.*³⁰ for further discussion).

One multi-validated dataset, the ‘filtered yeast interactome’ dataset⁸, contains only those interactions observed by at least two different methods and was generated with the express aim of reducing the number of false positives. Retaining only those interactions observed at least twice is an approach also adopted by Batada *et al.*⁷ to produce their ‘high-confidence’ (HC) dataset. This is currently the largest and most reliable set of yeast protein-protein interaction data. It combines the literature-curated data of Reguly *et al.*²⁵ with high-throughput data from a variety of sources, yielding a large number of interactions.

In their study, Batada and co-workers⁷ suggest a ‘new view’ of the yeast PPI network. They propose that the previously reported suppression of hub-hub interactions is an artifact^{10,28}, and that hubs are relatively homogenous, with no apparent distinction between the hitherto described ‘party’ and ‘date’ hubs⁸. They suggest that the main connected component of this network is much larger than previously appreciated, with relatively few disconnected components. However, they do not address the issues of sampling and biases associated with multiple validation.

Keeping only those data that can be validated by multiple observations will, by definition, remove interactions. The union of the five large datasets from which the HC set was drawn produces a list of 11,600 interactions between 4,500 proteins⁷. The final HC dataset has 9,300 interactions among 3,000 proteins. This means that not only is the known network a sample of the complete biological network because of the incompleteness of the experimentally identified interactions^{21,31}, but in addition the multi-validated dataset is a further subsample of all of the known interactions. Multiple validation is in some ways desirable, as it will decrease the false-positive rate. However, it introduces biases of its own: accepting only data that are observed at least twice results in a tendency to reject certain classes of data. For example, the set of interactions that will be retained will tend to be biased toward those that are highly studied. This has a drastic effect on network topology.

We demonstrate the alteration in topology by analysis of both the LC²⁵ and the HC⁷ data. In the LC network, the largest connected component includes 98% of the proteins, and there

are an additional 27 disconnected components (Fig. 1). If we apply the process of data validation such that interactions are included in the network only if they are observed in at least two studies, the largest connected component includes 79% of the proteins, and the number of disconnected components increases three-fold. If we further increase the stringency of validation, we further decrease the size of the largest component and increase the number of components. The HC dataset is somewhat different, as it already includes only those interactions that have been observed at least twice. However, we can again increase the stringency of validation, keeping only those interactions observed at least three times. As with the multiple validation of the LC dataset, the number of components increases, and the size of the largest component decreases.

Batada *et al.*⁷ have likened the yeast protein interaction network derived from the HC dataset to dense 'stratus'-type clouds, as compared to the clumpier 'altocumulus' form of previous datasets. As can be seen from our analysis (Fig. 1), this, too, is dependent on the stringency of validation, with the more stringently validated interactions giving rise to a clumpier, less connected network.

The topological changes evident in Figure 1 have a major effect on many of the global statistics that are commonly calculated for networks. Table 1 shows the behavior of several statistics for the HC and LC datasets. Note that none of these properties appears to be stable with respect to multiple validation, and most differ substantially between the HC and LC datasets. These differences are important, because they can change the biological conclusions drawn from the data.

Beyond global summary statistics, many other network properties are also sensitive to differences in data handling. For example, the LC data have a negative degree correlation between the degree of a node and the average degree of its neighbor; this has been suggested to support a hub-centric view of modularity¹⁰. However, this is only true of the raw data. If we again retain only those interactions observed twice or three times, the correlation first disappears, and then becomes positive (Fig. 2). Conclusions regarding modularity that are derived from the large-scale structure of the network are therefore surely suspect.

To be credible, any biological conclusions drawn from network structure should be robust with respect to the additional validation of interactions. However, we have demonstrated that the removal of interactions results in a radical change of many network properties. The effects of multiple validation will be similar in any dataset until we have

Table 1 Statistics for the LC²⁵ and the HC⁷ datasets, showing the effect of multiple validation on global network properties

Network	LC-1	LC-2	LC-3	HC-2	HC-3
Nodes	3,289	1,528	945	2,998	1,628
Edges	11,334	2,844	1,274	9,258	3,045
Connected components	28	90	124	101	200
Percentage nodes in largest component	98.0	79.4	36.3	91.8	55.7
Percentage edges in largest component	99.6	89.9	47.3	98.3	75.8
Mean degree	7.00	4.21	3.51	6.61	5.09
Diameter	12	19	19	15	30
Mean eccentricity	8.14	13.8	14.2	9.98	22.2
Radius	6	10	10	8	15
Mean shortest path length	4.22	7.54	7.56	4.90	9.99
Mean clustering coefficient	0.27	0.33	0.30	0.29	0.40

LC-1, LC-2 and LC-3 are literature-curated networks containing interactions observed in at least one, two or three different experimental studies, respectively. HC-2 and HC-3 contain high-confidence interactions observed at least twice and at least three times, respectively. Statistics are calculated for nodes in the largest connected component only. Degree, number of neighbors of a given node; diameter, length of the longest of all shortest paths between node pairs; eccentricity, length of the longest of all shortest paths between a given node and any other node; radius, minimum value of eccentricity over all nodes; clustering coefficient, proportion of possible edges between the neighbors of a given node that are actually present. Definitions and details of methods can be found in ref. 35.

unbiased multiple observations of all interactions.

Thus, the highest-quality datasets currently available contain interactions that, though *reliable*, are not necessarily *representative* of the network as a whole. This means that although any individual interaction has a high probability of being correct, drawing conclusions about the structure of the network as a whole is problematic.

Biological significance

Biases inherent in different protein interaction datasets can potentially be misleading when used to try to understand biology and,

if not controlled for, will be compounded in a multi-validated dataset. Thus, what is a topological analysis of these samples really telling us about the complete network²⁴, let alone biology? Biases in data can result in artifacts such as suppression of hub-hub interactions^{27,32}, and so it is necessary to be cautious in interpreting the biological meaning of node connectivity. This means that many conclusions about coarse-grained global measures, such as network topology, should be viewed with some skepticism.

With the current state of the data, it is important to be very cautious about inferring biological significance from large-scale

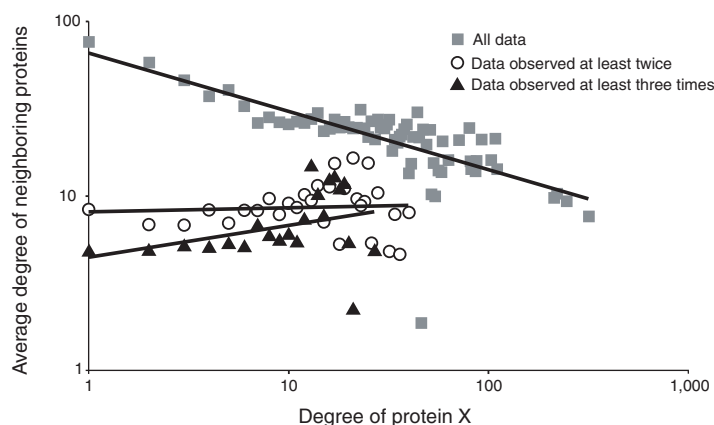


Figure 2 Topological analysis of the literature-curated dataset of Reguly *et al.*²⁵. This shows the connectivity of each protein versus its interacting partner(s) for the network composed of all LC interaction data (gray squares), interactions observed in at least two different experimental studies (open circles) and interactions observed in at least three different experimental studies (black triangles).

topological parameters. What does it mean that a network is scale free if any type of network can be turned into a scale-free one²³, or that a scale-free network loses this property²² when it is sampled? Can the proposed 'exponential core' at the heart of the PPI network²⁸ be assigned biological significance when it is likely that essential proteins are more likely to be highly studied (and therefore more extensively sampled)²⁷? Given that the correlation between nodes in the network and their neighbors¹⁰ can be radically altered through simple changes to data handling procedures, is it reasonable to speculate on the possibility that this correlation gives rise to robustness? Can we tell whether functional modularity arises easily from the structure of the network¹⁰ or depends on regulation of expression¹⁷? We argue that at this time it is impossible to tell.

Conclusions

Graph theory can be successfully and usefully applied to both human-made and naturally occurring networks. However, within a human-made network, such as the World Wide Web, every link is homogeneous and well defined. This means that it is possible to represent the structure of the complete network through visualizations and metrics derived from samples. In contrast, many of our existing models of biological networks are not representative of the underlying systems because of the varied nature of the interactions. For example, biological interactions vary in their nature and are spatially and temporally heterogeneous. As a result, although it is possible to create an abstract representation of these associations, the heterogeneity of the interactions necessitates appropriate sampling and analysis.

Abstractions are, of course, vital to understanding all sciences; this is particularly true in a field as complex as biology. Even so, oversimplifications can be positively misleading. Network views of PPIs are undoubtedly powerful when a detailed view of a given subsystem is analyzed³³. The power comes from reliable, complete, unbiased data that are understood in detail and make full use of dynamics, where appropriate. It is not even clear what a universal PPI network of a multicellular organism represents, given that many interactions are developmentally related or tissue specific. However, even if it proves possible to describe complex PPI networks in a manner similar to physical systems, it is vital to understand data quality, origins and biases, or we are likely to be misled.

To make reliable inferences, we must have reasonably complete datasets that accurately represent the interactions in the proteome. If biases exist, their effects on network properties need to be understood and their influence factored into any analysis. Detailed information concerning the nature of interactions is required, including the specific functional implications of an interaction. Only in this way can we ensure that network analysis and biological understanding are connected.

ACKNOWLEDGMENTS

L.H. was supported by Wellcome Trust VIP funding to S.C.L. J.W.P. is supported by a Biotechnology and Biological Sciences Research Council project grant (BB/C515412/1) to D.L.R. We thank K. Hentges for critical reading of the manuscript.

1. Uetz, P. *et al.* *Nature* **403**, 623–627 (2000).
2. Ito, T. *et al.* *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
3. Gavin, A.C. *et al.* *Nature* **440**, 631–636 (2006).
4. Krogan, N.J. *et al.* *Nature* **440**, 637–643 (2006).

5. Barabasi, A.L. & Albert, R. *Science* **286**, 509–512 (1999).
6. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. *Nature* **411**, 41–42 (2001).
7. Batada, N.N. *et al.* *PLoS Biol.* **4**, e317 (2006).
8. Han, J.D. *et al.* *Nature* **430**, 88–93 (2004).
9. Przulj, N., Wagle, D.A. & Jurisica, I. *Bioinformatics* **20**, 340–348 (2004).
10. Maslov, S. & Sneppen, K. *Science* **296**, 910–913 (2002).
11. Albert, R., Jeong, H. & Barabasi, A.L. *Nature* **406**, 378–382 (2000).
12. Goh, K.I., Oh, E., Jeong, H., Kahng, B. & Kim, D. *Proc. Natl. Acad. Sci. USA* **99**, 12583–12588 (2002).
13. Berg, J., Lässig, M. & Wagner, A. *BMC Evol. Biol.* **4**, 51 (2004).
14. Salathe, M., May, R.M. & Bonhoeffer, S. *J. R. Soc. Interface* **2**, 533–536 (2005).
15. Ispolatov, I., Krapiivsky, P.L. & Yuryev, A. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **71**, 061911 (2005).
16. Bersini, H., Lenaerts, T. & Santos, F.C. *J. Theor. Biol.* **241**, 488–505 (2006).
17. Batada, N.N., Hurst, L.D. & Tyers, M. *PLoS Comput. Biol.* **2**, e88 (2006).
18. Bertin, N. *et al.* *PLoS Biol.* **5**, e153 (2007).
19. Batada, N.N. *et al.* *PLoS Biol.* **5**, e154 (2007).
20. von Mering, C. *et al.* *Nature* **417**, 399–403 (2002).
21. Hart, G.T., Ramani, A.K. & Marcotte, E.M. *Genome Biol.* **7**, 120 (2006).
22. Stumpf, M.P., Wiuf, C. & May, R.M. *Proc. Natl. Acad. Sci. USA* **102**, 4221–4224 (2005).
23. Han, J.D., Dupuy, D., Bertin, N., Cusick, M.E. & Vidal, M. *Nat. Biotechnol.* **23**, 839–844 (2005).
24. de Silva, E. *et al.* *BMC Biol.* **4**, 39 (2006).
25. Regulý, T. *et al.* *J. Biol.* **5**, 11 (2006).
26. Ekman, D., Light, S., Bjorklund, A.K. & Elofsson, A. *Genome Biol.* **7**, R45 (2006).
27. Hakes, L., Robertson, D.L. & Oliver, S.G. *BMC Genomics* **6**, 131 (2005).
28. Pereira-Leal, J.B., Audit, B., Peregrin-Alvarez, J.M. & Ouzounis, C.A. *Mol. Biol. Evol.* **22**, 421–425 (2005).
29. Barabasi, A.L. & Oltvai, Z.N. *Nat. Rev. Genet.* **5**, 101–113 (2004).
30. Suthram, S., Shlomi, T., Ruppin, E., Sharan, R. & Ideker, T. *BMC Bioinformatics* **7**, 360 (2006).
31. Stumpf, M.P. & Wiuf, C. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **72**, 036118 (2005).
32. Aloy, P. & Russell, R.B. *FEBS Lett.* **530**, 253–254 (2002).
33. Xia, K. *et al.* *PLoS Comput. Biol.* **2**, e145 (2006).
34. Adai, A.T., Date, S.V., Wieland, S. & Marcotte, E.M. *J. Mol. Biol.* **340**, 179–190 (2004).
35. Brandes, U. & Erlebach, T. *Network Analysis* (Springer, Berlin, Germany, 2005).