JOURNAL OF THE ROYAL SOCIETY Interface

Modelling protein–protein interaction networks via a stickiness index

Nataaa Pr~ulj and Desmond J Higham

J. R. Soc. Interface 2006 **3**, 711-716
doi: 10.1098/rsif.2006.0147

JOURNAL OF THE ROYAL SOCIETY Interface
J. R. Soc. Interface (2006) **3**, 711–716
doi:10.1098/rsif.2006.0147
Published online 22 August 2006

REPORT

Modelling protein–protein interaction networks via a stickiness index

Nataša Pržulj[1,*] and Desmond J. Higham[2]

[1]*Department of Computer Science, University of California, Irvine, CA 92697-3425, USA*
[2]*Department of Mathematics, University of Strathclyde, Glasgow G11XH, UK*

What type of connectivity structure are we seeing in protein–protein interaction networks? A number of random graph models have been mooted. After fitting model parameters to real data, the models can be judged by their success in reproducing key network properties. Here, we propose a very simple random graph model that inserts a connection according to the degree, or 'stickiness', of the two proteins involved. This model can be regarded as a testable distillation of more sophisticated versions that attempt to account for the presence of interaction surfaces or binding domains. By computing a range of network similarity measures, including relative graphlet frequency distance, we find that our model outperforms other random graph classes. In particular, we show that given the underlying degree information, fitting a stickiness model produces better results than simply choosing a degree-matching graph uniformly at random. Therefore, the results lend support to the basic modelling methodology.

Keywords: protein–protein interaction networks; network models; network properties

1. INTRODUCTION AND MODEL

A protein–protein interaction (PPI) network is commonly viewed as an unweighted, undirected graph. Each node in the graph represents a protein and an edge between a pair of nodes indicates that these proteins have been observed to interact physically (Ito *et al.* 2000; Uetz *et al.* 2000; Giot *et al.* 2003; Li *et al.* 2004; Rual *et al.* 2005; Stelzl *et al.* 2005). The types of connectivity patterns that arise are neither completely random, in the classical Erdös–Rényi sense (henceforth denoted by 'ER'), nor completely deterministic (Grindrod & Kibble 2004).

In an attempt to understand and describe the PPI connectivities, a number of models, i.e. formulae for generating edges in some probabilistic sense, have been proposed and tested against observed networks (Jeong *et al.* 2001; Maslov & Sneppen, 2002; Barabási *et al.* 2003; Pržulj *et al.* 2004; de Silva & Stumpf 2005). Many works have focused on matching degree distributions and recovering a scale-free law (Jeong *et al.* 2001; Maslov & Sneppen 2002; Barabási *et al.* 2003; Salathé *et al.* 2005), although whether PPI networks are really scale-free is still the subject of debate (Pržulj *et al.* 2004; Han *et al.* 2005; Dupuy *et al.* 2006; Friedel & Zimmer 2006; Khanin & Wit 2006). Our aim here is to present a new, pared-down, but biologically motivated model that simplifies previous work to the extent that fitting parameters and comparing local and global graph properties become meaningful and revealing.

Among the few existing models that incorporate some biological justification are those of Caldarelli *et al.* (2002), Thomas *et al.* (2003) and Deeds *et al.* (2006). These related models have in common the idea that proteins interact because they share complimentary physical aspects, a concept that is consistent with the underlying biochemistry. Following Thomas *et al.* (2003), we will refer to these physical aspects as binding domains. The approach in these papers is to generate graphs by assigning binding domain information to the nodes at random and then inserting links probabilistically according to some pairwise matching criterion. The aim is then to reproduce properties observed in real PPI networks, most notably the degree distribution. We also mention that a refined 'lock-and-key' version of the model from Thomas *et al.* (2003) has been used to extract protein-level detail from real datasets (Morrison *et al.* 2006), further justifying the modelling approach.

Presently, it would be a very challenging task to infer the number and distribution of distinct binding domains from a real PPI network (Bateman 2002; Deng *et al.* 2002), not least because the networks are known to be noisy (Sprinzak *et al.* 2003). For this reason, it is difficult to decide whether the models from (Caldarelli *et al.* 2002; Thomas *et al.* 2003; Deeds *et al.* 2006) are being tested under realistic parameter ranges. Therefore, we propose a simplified model that attempts to summarize the abundance and popularity of binding domains on a protein as a single number based on its normalized degree; we call this number the *stickiness index*. The model has the benefit of being tunable to the given degree structure of a PPI network. In this way, a benchmark model that captures the essence of Caldarelli *et al.* (2002), Deeds *et al.* (2006), Thomas *et al.* (2003) can be tested.

Our work can be motivated by two main assumptions.

Assumption 1. Having a high degree implies that a protein has many binding domains and/or its binding domains are commonly involved in interactions.

Assumption 2. A pair of proteins is more likely to interact (share complementary binding domains) if both have high stickiness indices, and correspondingly less likely to interact if one or both have a low stickiness index. Thus, we take the product of the two stickiness indices to define the probability of interaction—this borrows from the concept of an AND gate in Boolean logic (Ben-Ari 2001) and the idea of a rank-one approximation in dimension reduction (Eldén 2006).

*Author for correspondence (natasha@ics.uci.edu).

Received 21 June 2006
Accepted 14 July 2006

The following pseudocode defines our model.

input $\{\deg_i\}_{i=1}^N$, list of degrees of N nodes
output $\{w_{ij}\}_{i,j=1}^N$, adjacency matrix from model

```
for i=1 to N
    θ_i = deg_i / sqrt(Σ_{j=1}^N deg_j)
end
Initialize all w_ij=0
for i=1 to N
    for j=1 to N
        compute a uniform (0, 1) sample, r
        if r≤θ_iθ_j
            w_ij=1 and w_ji=1
        end if
    end for
end for
```

This choice of stickiness index θ_i ensures that the ith node in the model has the *expected* degree $\deg_i$. Moreover, under assumption 2, this definition of stickiness in terms of degree is the only one that captures the correct expected degree. Details are given in appendix A.

Our stickiness index coincides with the concept of *fitness* in Caldarelli *et al.* (2002), with a notable distinction that fitness in Caldarelli *et al.* (2002) is assigned at random, with a focus on the resulting degree distribution, whereas stickiness above is assigned deterministically, based on the unique choice that matches the expected degrees. Since we do not require any other parameter fitting, this approach allows us to perform a 'proof of principle' test of the basic idea that links can be modelled via mutual compatibility.

Note that high-degree proteins in the present PPI networks may not necessarily contain a plenty of binding domains, as implied by our assumption 1. Instead, their high connectivities may be artefacts of *technical false positives*, auto-activators or 'sticky' proteins, or owing to *biological false positives*, as some PPIs can occur in the experimental procedure, but not *in vivo* because protein pairs are not expressed at the same time, in the same sub-cellular compartment, or in the same tissue (Han *et al.* 2005). Thus, our assumption 1 may be a severe oversimplification for some proteins in the present PPI datasets. Nevertheless, as PPI detection biotechnologies improve to produce cleaner, higher-confidence PPI data, assumption 1 will become more descriptive of the observed networks.

A multitude of random graph models that reproduce scale-free degree distributions have been proposed, although the relevance of scale-freeness to PPI networks has been questioned (Pržulj *et al.* 2004; Han *et al.* 2005; Dupuy *et al.* 2006; Friedel & Zimmer 2006; Khanin & Wit 2006). The most notable such models are those based on biologically motivated *gene duplication and mutation* network growth principles (Vazquez *et al.* 2001; Pastor-Satorras *et al.* 2003; Wagner 2003; Goh *et al.* 2004). In these models, networks grow by duplication of nodes (genes), and as a node gets duplicated, it inherits most of the neighbours (interactions) of the parent node, but gains some new neighbours as well. Thus, a hybrid model having properties of both the gene duplication–mutation model and the stickiness index-based model is a promising future direction. In such a model, a duplicated gene would inherit the parent's stickiness index along with many of the parent's neighbours, as in a gene duplication–mutation model and it would gain new neighbours in proportion to its inherited stickiness index and stickiness indices of the nodes already in the network, as in our stickiness index-based model.

We remark that early tests on low confidence data in (Maslov & Sneppen 2002) suggest that PPI networks have a bias against connections between high-degree proteins. This is potentially at odds with the models in Caldarelli *et al.* (2002), Deeds *et al.* (2006), Thomas *et al.* (2003), where sets of proteins that share matching and commonly occurring (high fitness) physical aspects will interact and all have high degree. In our simple model, we assign edges independently, but it would be possible to add a post-processing stage in which the links were rewired in order to test various types of correlation. Hence, a further application of our model is in studying correlation effects in PPI network topology.

2. EXPERIMENTS AND RESULTS

Comparing large real-world networks is computationally intensive as it involves an NP-complete *subgraph isomorphism problem* (West 2001). Thus, simple heuristics measuring *global* and *local* network properties have been used. The most commonly examined global network properties are the *degree distribution*, *clustering coefficient* and network *diameter* (see Newman (2003) for a detailed survey). More recently, bottom-up local approaches to study a network structure have been proposed (Milo *et al.* 2002; Shen-Orr *et al.* 2002; Pržulj *et al.* 2004). Analogous to sequence motifs, *network motifs* have been defined as subgraphs that recur in a network at frequencies much higher than those found in randomized networks (Milo *et al.* 2002; Shen-Orr *et al.* 2002; Milo *et al.* 2004); they were used to uncover basic functional units in various real-world networks. To account for frequencies of occurrence of all small subgraphs rather than for only the over-represented ones, *graphlets* were defined as small connected non-isomorphic induced subgraphs of a large network and their *relative frequencies* were used to define a new *distance* measure between two networks (Pržulj *et al.* 2004).

To examine the fit of our new stickiness index-based model of PPI networks, we use all these standard global and local network parameters. The relative graphlet frequency distance is the most demanding network similarity measure, imposing 29 different constraints on the networks being compared (details in Pržulj *et al.* 2004); hence, we use it as our main comparison tool. We compared 14 large publicly available PPI networks with sample networks from five models, including the stickiness model.

We used PPI networks of the following eukaryotic organisms: yeast *Saccharomyces cerevisiae*; fruitfly *Drosophila melanogaster*; nematode worm *Caenorhabditis elegans*; and human. Several different datasets are

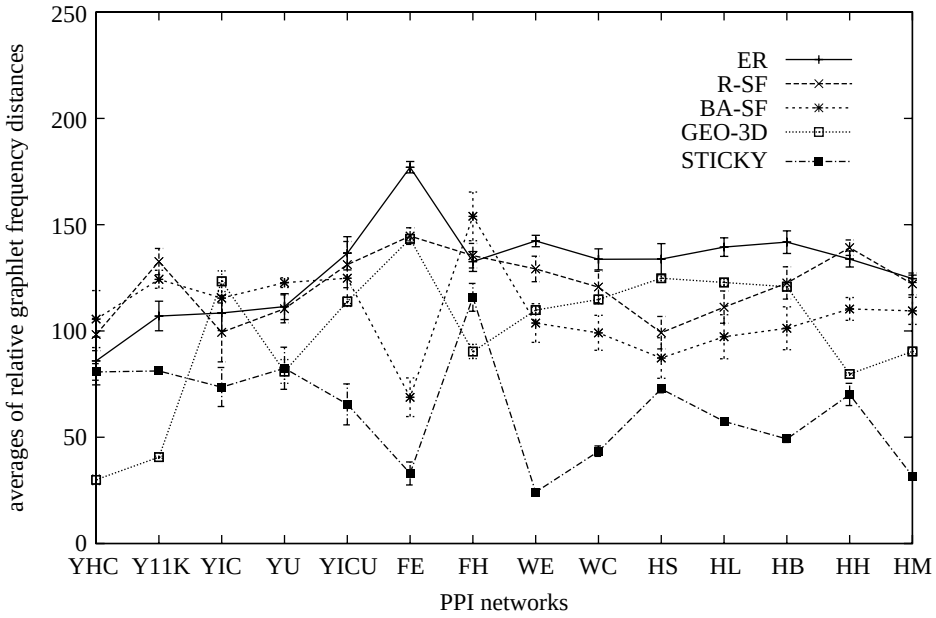

Figure 1. Relative graphlet frequency distances (y-axis) between the 14 PPI networks (x-axis) and their corresponding model networks. The lower the number, the better the fit. Averages of distances between 25 sample networks and the corresponding PPI network are presented for each random graph model and each PPI network. Points are joined only for clarity. The error bar around a point spans one standard deviation above and below (in some cases, error bars are barely visible, since they are of the size of the point). Labels on the horizontal axis are described in the text.

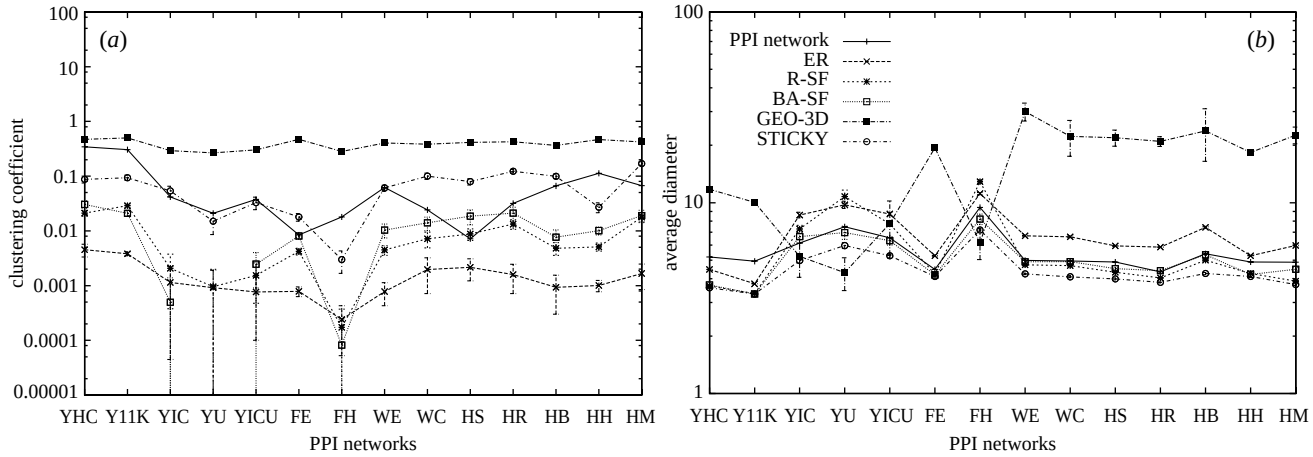

Figure 2. (a) Clustering coefficients of 14 PPI networks and averages of clustering coefficients of 25 model networks corresponding to a PPI network. (b) Average diameters of the 14 PPI networks and averages of average diameters of 25 model networks corresponding to a PPI network. Error bars and labels are as described in the legend of figure 1.

available for yeast and human, so we analysed five yeast PPI networks of different confidence levels obtained from three different high-throughput studies (Ito *et al.* 2000; Uetz *et al.* 2000; von Mering *et al.* 2002), as well as five human PPI networks obtained from the two recent high-throughput studies (Rual *et al.* 2005; Stelzl *et al.* 2005) and three curated databases (Zanzoni *et al.* 2002; Bader *et al.* 2003; Peri *et al.* 2004). We denote by 'YHC' the high-confidence yeast PPI network from von Mering *et al.* (2002), by 'Y11K' the yeast PPI network defined by the top 11 000 interactions in the von Mering *et al.* classification (von Mering *et al.* 2002), by 'YIC' the Ito *et al.* 'core' yeast PPI network (Ito *et al.* 2000), by 'YU' the Uetz *et al.* yeast PPI network (Uetz *et al.* 2000), and by 'YICU' the union of YIC and YU yeast PPI networks (we combined them as in (Han *et al.* 2005) to increase coverage). 'FE' and 'FH' denote the fruitfly *D. melanogaster* entire and high-confidence PPI networks from (Giot *et al.* 2003). Similarly, 'WE' and 'WC' denote the worm *C. elegans* entire and 'core' PPI networks from (Li *et al.* 2004). Finally, 'HS', 'HR', 'HB', 'HH' and 'HM' stand for human PPI networks from yeast two-hybrid (Y2H) screens by Stelzl *et al.* (Stelzl *et al.* 2005) and Rual *et al.* (Rual *et al.* 2005) and from curated databases BIND (Bader *et al.* 2003), HPRD (Peri *et al.* 2004) and MINT (Zanzoni *et al.* 2002), respectively, (BIND, HPRD and MINT data were downloaded from OPHID (Brown & Jurisica 2005) on 10 February 2006). Note that YHC and Y11K networks are mainly coming from tandem affinity purifications (Gavin *et al.* 2002) and high-throughput

mass spectrometric protein complex identification (Ho *et al.* 2002), while YIC, YU, YICU, FE, FH, WE, WH, HS and HR are yeast two-hybrid, and HB, HH, and HM are a result of human curation (BIND, HPRD and MINT). Thus, we are using PPI networks of different confidence levels that come from a range of high-throughput PPI detection biotechnologies as well as from human curation.

We compared these PPI networks with the following five model networks: ER random graphs (Erdös & Rényi 1959, 1960); random graphs with exactly the same degree distribution as that of a PPI network (Bender & Canfield 1978; Newman 2002) (denoted 'R-SF' for 'random scale-free'); Barabasi–Albert scale-free networks (Barabási & Albert 1999) (denoted by 'BA-SF'); three-dimensional geometric random graphs (Penrose 2003) (denoted by 'GEO-3D'); and the stickiness model networks described previously (denoted by 'STICKY').

For each of the 14 PPI networks and for each of the five models, we compared the PPI network with 25 samples from the model. Each sample matched the number of nodes and edges in the corresponding PPI network.

Average relative graphlet frequency distances between the PPI and the corresponding model networks for each of the five network models are presented in figure 1. The stickiness model shows an improved fit over all other network models with respect to relative graphlet frequency distances in 10 out of 14 tested PPI networks (filled squares in figure 1); it fits as well as the GEO-3D model (open squares in figure 1) in one and is outperformed by the GEO-3D model in three PPI networks. In addition, this model reproduces global network properties, such as the degree distribution (see appendix A), the clustering coefficients (open circles in figure 2*a*) and the average diameters of PPI networks (open circles in figure 2*b*).

It is of particular note that the R-SF model does not perform as well as the stickiness model. This means that, given the degree distribution of a PPI network,

(i) simply drawing a network uniformly at random from the class of all networks that match the degree distribution is less successful in capturing the underlying substructure than
(ii) enhancing this degree information by using the simple modelling insights summarized in assumptions 1 and 2.

3. CONCLUSIONS

Overall, the stickiness framework produces a convenient, parameter-free random network that is motivated by transparent modelling arguments and may be regarded as a simplified, testable distillation of more sophisticated models. The results give further justification for the modelling approaches in (Caldarelli *et al.* 2002; Thomas *et al.* 2003; Deeds *et al.* 2006). Since the model accurately reproduces all widely used quantitative measures, it also provides a benchmark against which others may be compared.

We thank the referees for their valuable feedback.

APPENDIX A

Suppose $A \in \mathbb{R}^{N\times N}$ is the PPI network adjacency matrix, then $a_{ij}=a_{ji}=1$ if proteins i and j are connected and $a_{ij}=a_{ji}=0$ otherwise. We are using $\deg_i := \sum_{j=1}^{N} a_{ij}$ to denote the degree of protein i.

Suppose that some function of the degree, $f^{[i]}(\deg_i)$, defines the stickiness index of protein i, then under assumption 2 (and independently for each distinct pair of proteins),

$$\mathbb{P}(i \leftrightarrow j) = f^{[i]}(\deg_i)\cdot f^{[j]}(\deg_j),$$

where $i \leftrightarrow j$ denotes the event that i and j are connected.

In order to match the PPI network degree with the expected degree from the model, we require

$$\begin{aligned}\deg_i &= \mathbb{E}[\text{degree of node } i \text{ in model}]\\ &= \sum_{j=1}^{N}\mathbb{P}(i \leftrightarrow j)\\ &= \sum_{j=1}^{N} f^{[i]}(\deg_i)\cdot f^{[j]}(\deg_j)\\ &= f^{[i]}(\deg_i)\sum_{j=1}^{N} f^{[j]}(\deg_j).\end{aligned}$$

Let $C=\sum_{j=1}^{N} f^{[j]}(\deg_j)$. Then, the formula above tells us that $\deg_i = C f^{[i]}(\deg_i)$, and thus

$$f^{[i]}(\deg_i) = \frac{\deg_i}{C}.$$

Summing over i shows that $C^2=\sum_{i=1}^{N}\deg_i$. We conclude that

$$f^{[i]}(\deg_i) = \frac{\deg_i}{\sqrt{\sum_{j=1}^{N}\deg_j}},$$

confirming that our stickiness index, θ_i, is uniquely defined under our assumptions.

We note that for all probabilities to be in the range $[0, 1]$, we require $\theta_i\theta_j \le 1$ for all i,j. (Assuming that all proteins have at least one interaction, a sufficient condition is that the product of the two largest degrees is bounded by N.) This property holds for all networks considered here.

As discussed in (Caldarelli *et al.* 2002), an intuitively reasonable alternative to the multiplicative model is the additive version

$$\mathbb{P}(i \leftrightarrow j) = g^{[i]}(\deg_i) + g^{[j]}(\deg_j).$$

However, copying the same style of analysis leads to the conclusion that

$$g^{[i]}(\deg_i) = \frac{\deg_i}{N} - \frac{1}{2N}\sum_{k=1}^{N}\deg_k,$$

so that

$$\mathbb{P}(i \leftrightarrow j) = \frac{1}{N}\left(\deg_i + \deg_j - \frac{1}{N}\sum_{k=1}^{N}\deg_k\right).$$

Since many proteins have degree less than half the network average, this model breaks down owing to the assignment of negative probabilities.

REFERENCES

Bader, G. D., Betel, D. & Hogue, C. W. V. 2003 BIND: the biomolecular interaction network database. *Nucleic Acids Res.* **31**, 248–250. (doi:10.1093/nar/gkg056)

Barabási, A.-L. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–12. (doi:10.1126/science.286.5439.509)

Barabási, A.-L., Dezso, Z., Ravasz, E., Yook, Z.-H. & Oltvai, Z. N. 2003 Scale-free and hierarchical structures in complex networks. In *Modeling of complex systems: Seventh Granada Lectures. AIP Conference Proceedings*, vol. 661, p. 1–16. College Park, MA: AIP.

Bateman, A. *et al.* 2002 The pfam protein families database. *Nucleic Acids Res.* **30**, 276–280. (doi:10.1093/nar/30.1.276)

Ben-Ari, M. 2001 *Mathematical logic for computer science.* Berlin, Germany: Springer.

Bender, E. A. & Canfield, E. R. 1978 The asymptotic number of labeled graphs with given degree sequences. *J. Comb. Theor. A* **24**, 296–307. (doi:10.1016/0097-3165(78)90059-6)

Brown, K. & Jurisica, I. 2005 Online predicted human interaction database. *Bioinformatics* **21**, 2076–2082.

Caldarelli, G., Capocci, A., De Los Rios, P. & Munoz, M. A. 2002 Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **89**, 258702-1-4. (doi:10.1103/PhysRevLett.89.258702)

de Silva, E. & Stumpf, M. P. H. 2005 Complex networks and simple models in biology. *J. R. Soc. Interface* **2**, 419–430. (doi:10.1098/rsif.2005.0067)

Deeds, Eric J., Ashenberg, Orr & Shakhnovich, Eugene I. 2006 A simple physical model for scaling in protein–protein interaction networks. *Proc. Natl Acad. Sci.* **103**, 311–316. (doi:10.1073/pnas.0509715102)

Deng, M., Mehta, S., Sun, F. & Chen, T. 2002 Inferring domain–domain interactions from protein–protein interactions. *Genome Res.* **12**, 1540–1548. (doi:10.1101/gr.153002)

Dupuy, D., Bertin, N., Cusick, M. E., Han, J.-D. J. & Vidal, M. 2006 Reply to toward the complete interactome. *Nat. Biotechnol.* **24**, 615–615. (doi:10.1038/nbt0606-615a)

Eldén, L. 2006 *Matrix methods in data mining and pattern recognition.* SIAM, PA.

Erdös, P. & Rényi, A. 1959 On random graphs. *Publicationes Mathematicae* **6**, 290–297.

Erdös, P. & Rényi, A. 1960 On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61.

Friedel, C. C. & Zimmer, R. 2006 Toward the complete interactome. *Nat. Biotechnol.* **24**, 614–615. (doi:10.1038/nbt0606-614)

Gavin, A. C. *et al.* 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147. (doi:10.1038/415141a)

Giot, L. *et al.* 2003 A protein interaction map of drosophila melanogaster. *Science* **302**, 1727–1736. (doi:10.1126/science.1090289)

Goh, K.-I., Kahng, B. & Kim, D. 2004 Hybrid network model: the protein and the protein family interaction networks. See http://arxiv.org/abs/q-bio.MN/0312009.

Grindrod, P. & Kibble, M. 2004 Review of uses of network and graph theory concepts within proteomics. *Expert Rev. Proteomics* **1**, 89–98. (doi:10.1586/14789450.1.2.229)

Han, J. D. H., Dupuy, D., Bertin, N., Cusick, M. E. & Vidal, M. 2005 Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.* **23**, 839–844. (doi:10.1038/nbt1116)

Ho, Y. *et al.* 2002 Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180–183. (doi:10.1038/415180a)

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. & Sakaki, Y. 2000 Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* **97**, 1143–1147. (doi:10.1073/pnas.97.3.1143)

Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. 2001 Lethality and centrality in protein networks. *Nature* **411**, 41–42. (doi:10.1038/35075138)

Khanin, R. & Wit, F. 2006 How scale-free are gene networks? *J. Comput. Biol.* **13**, 810–818. (doi:10.1089/cmb.2006.13.810)

Li, S. *et al.* 2004 A map of the interactome network of the metazoan *c. elegans. Science* **303**, 540–543. (doi:10.1126/science.1091403)

Maslov, S. & Sneppen, K. 2002 Specificity and stability in topology of protein networks. *Science* **296**, 910–913. (doi:10.1126/science.1065103)

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. 2004 Superfamilies of evolved and designed networks. *Science* **303**, 1538–1542. (doi:10.1126/science.1089167)

Milo, R., Shen-Orr, S. S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. 2002 Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827. (doi:10.1126/science.298.5594.824)

Morrison, J. L., Breitling, R., Higham, D. J. & Gilbert, D. R. 2006 A lock-and-key model for protein–protein interactions. *Bioinformatics* **22**, 2010–2019. (doi:10.1093/bioinformatics/btl338)

Newman, M. E. J. 2002 Random graphs as models of networks. In *Handbook of graphs and networks* (ed. S. Bornholdt & H. G. Schuster). Berlin, Germany: Wiley-VHC.

Newman, M. E. J. 2003 The structure and function of complex networks. *SIAM Rev.* **45**, 167–256. (doi:10.1137/S0036144450342480)

Pastor-Satorras, R., Smith, E. & Sole, V. 2003 Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **222**, 199–210. (doi:10.1016/S0022-5193(03)00028-6)

Penrose, M. 2003 *Geometric random graphs.* Oxford, UK: Oxford Univeristy Press.

Peri, S. *et al.* 2004 Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **32**, 1362–4962. Database issue:D497-501. (doi:10.1093/nar/gkh070)

Pržulj, N., Corneil, D. G. & Jurisica, I. 2004 Modeling interactome: scale-free or geometric? *Bioinformatics* **20**, 3508–3515. (doi:10.1093/bioinformatics/bth436)

Rual, J.-F. *et al.* 2005 Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–178. (doi:10.1038/nature04209)

Salathé, M., May, R. M. & Bonhoeffer, S. 2005 The evolution of network topology by selective removal. *J. R. Soc. Interface* **2**, 533–536.

Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. 2002 Network motifs in the transcriptional regulation network of *Escherichia coli. Nat. Genet.* **31**, 64–68. (doi:10.1038/ng881)

Sprinzak, E., Sattath, S. & Margalit, H. 2003 How reliable are experimental protein–protein interaction data? *J. Mol. Biol.* **327**, 919–923. (doi:10.1016/S0022-2836(03)00239-0)

Stelzl, U. *et al.* 2005 A human protein–protein interaction network: A resource for annotating the proteome. *Cell* **122**, 957–968. (doi:10.1016/j.cell.2005.08.029)

Thomas, A., Cannings, R., Monk, N. A. M. & Cannings, C. 2003 On the structure of protein–protein interaction networks. *Biochem. Soc. Trans.* **31**, 1491–1496.

Uetz, P. *et al.* 2000 A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627. (doi:10.1038/35001009)

Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. 2001 Modeling of protein interaction networks. *ComPlexUs* **1**, 38–44.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. 2002 Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403. (doi:10.1038/nature750)

Wagner, A. 2003 How the global structure of protein interaction networks evolves. *Proc. R. Soc. B* **270**, 457–466. (doi:10.1098/rspb.2002.2269)

West, D. B. 2001 *Introduction to graph theory*, 2nd edn. Upper Saddle River, NJ: Prentice Hall.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. & Cesareni, G. 2002 Mint: a molecular interaction database. *FEBS Letters* **513**, 135–140. (doi:10.1016/S0014-5793(01)03293-8)