JOURNAL OF THE ROYAL SOCIETY Interface

Complex networks and simple models in biology

Eric de Silva and Michael P.H Stumpf

J. R. Soc. Interface 2005 **2**, 419-430
doi: 10.1098/rsif.2005.0067

JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2005) **2**, 419–430
doi:10.1098/rsif.2005.0067
Published online 24 August 2005

REVIEW

Complex networks and simple models in biology

Eric de Silva and Michael P. H. Stumpf[†]

Theoretical Genomics Group, Division of Molecular Biosciences, Imperial Collage London, Wolfson Building, South Kensington Campus, London SW7 2AZ, UK

The analysis of molecular networks, such as transcriptional, metabolic and protein interaction networks, has progressed substantially because of the power of models from statistical physics. Increasingly, the data are becoming so detailed—though not always complete or correct—that the simple models are reaching the limits of their usefulness. Here, we will discuss how network information can be described and to some extent quantified. In particular statistics offers a range of tools, such as model selection, which have not yet been widely applied in the analysis of biological networks. We will also outline a number of present challenges posed by biological network data in systems biology, and the extent to which these can be addressed by new developments in statistics, physics and applied mathematics.

Keywords: biological networks; network models; network sampling; protein interactions; systems biology

1. INTRODUCTION

Following the enormous advances in functional genomics and molecular biology it is now possible to at least contemplate studying cellular processes at the level of a whole cell, rather than in isolation. Molecular networks, such as protein interaction (Uetz *et al.* 2000; Maslov & Sneppen 2002; Agrafioti *et al.* 2005), metabolic (Ma & Zeng 2003) and gene regulation networks (Ronen *et al.* 2002; Evangelisti & Wagner 2004) aim to capture such sets of biological processes in a single and coherent framework. In reality, of course, these different networks are intricately connected and interwoven inside a cell; protein products will interact with each other, regulate the expression of genes as well as digesting nutrients and catalysing basic biochemical reactions in a cells metabolism. We are still far away from being able to consolidate these different networks into a realistic *in-silico* organism.

The analysis and interpretation of present network data is, however, already challenging enough. Since the late 1990s research has been aided considerably by the work of a host of physicists (see Albert & Barabasi 2002; Dorogovtsev & Mendes 2003; Newman 2003*a*; Evans 2004, for mainly physics-oriented reviews). While the models proposed have, despite their elegant simplicity, been able to explain certain aspects of complex biological networks, they increasingly reach the limit of their usefulness given the amount of data becoming available. New models, based on sound statistical principles, and informed by bioinformatics are now slowly taking their place.

The theoretical underpinnings for the analysis of networks come from statistical physics, mathematics (in particular random graph theory; Bollobás 1998) and computer science. Despite several attempts over the last few years to apply concepts from these disciplines to biological networks, success has often been modest. There are numerous examples where terminology or concepts have been taken from, e.g. statistical physics, and wrongly applied in the description of molecular networks. We will first discuss how networks can be analysed statistically and described theoretically. The statistical analysis may either focus on (structural) properties of the network itself, or on biological properties of the constituents of the network. While the former is well advanced the latter will (even in the presence of high quality data) pose a range of fascinating and challenging problems.

Below we will first introduce the biological networks that are currently attracting most interest as framework for systems biology. After that we will discuss the different theoretical models that have been used to model complex biological (among many other types) networks before introducing a set of statistical tools and, more interestingly, problems. Whenever biological examples are discussed in this exposition they will have a distinctly evolutionary perspective.

2. BIOLOGICAL NETWORK DATA

As already mentioned we can, very coarsely, distinguish between three types of molecular networks.

[†]Author for correspondence (m.stumpf@imperial.ac.uk).

Received 3 June 2005
Accepted 14 July 2005

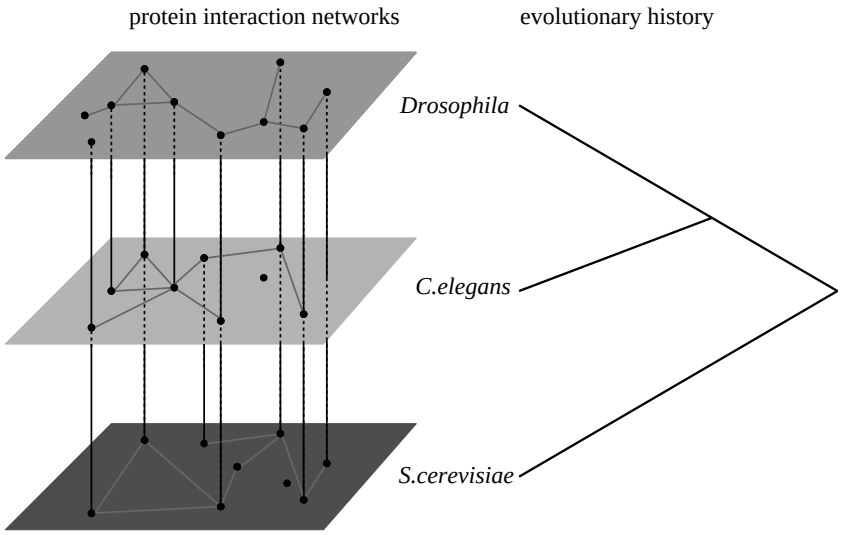

Figure 1. Protein interaction network data is collected in different organisms. Orthologous proteins are indicated by vertical lines, interactions between proteins by lines within the planes. Individual proteins are also related through their joined phylogeny.

Metabolic networks: these aim to describe the basic biochemistry in a cell. Biologically important reactions have been described in terms of reaction pathways and metabolic networks are systematic collections of such biochemical data.

Transcriptional networks: these consist of genes and a directed edge is added between two genes if one regulates the transcription of the other gene.

Protein interaction networks (PIN): an undirected edge is drawn between each pair of proteins for which there is evidence of a physical or biochemical interaction.

Making these distinctions and simplifications must necessarily neglect details of the biological processes. In reality these networks will be highly and intricately interconnected and factorizing them into distinct networks will ultimately underestimate the biological complexity.

These problems are exacerbated when one considers the often woeful quality of the data: for PIN the rates for false-positive and false-negative results are estimated to be around 40% (Bader *et al.* 2004; Tong *et al.* 2004). Bioinformatics and statistics may help to clean the data to some extent but improvements in the experimental techniques offer the only real solution to this problem. Although important and interesting (Lappe & Holm 2004) we will here not be concerned with such issues of quality control. Rather we will discuss what should be included in theoretical descriptions of complex networks in a biological setting.

It has to be kept in mind, though, that present network data are highly averaged and artificial constructs: the language of graph theory may simply be too static to usefully describe complex biological networks. We may in approximation seek to understand networks as entities that change over three different time-scales: (i) they will change over evolutionary time-scales between species (millions of years); (ii) they will change during the course of an organism's development (years); and finally, (iii) connections will be formed and lost in response to physiological change and external stimuli (sub-second to minutes). For PIN experimental methods can at the moment only resolve the changes in PIN structure accumulated between species (Fraser *et al.* 2002; Jordan *et al.* 2003; Qin *et al.* 2003), but data are not yet sufficiently reliable to make meaningful comparisons.

One of the fundamental evolutionary questions underlying comparative genomics and the fledgeling discipline of systems biology is illustrated in figure 1. Within each species' PIN, interactions introduce a dependence between interacting proteins, i.e. it may no longer be possible to consider them independently (Agrafioti *et al.* 2005). The phylogeny underlying the different model organisms introduces a further level of correlation (Li 1997; Felsenstein 2003). In the functional analysis of networks we will often have to include both types of correlation, which makes the correct statistical analysis of biological network data highly non-trivial.

Below we will first briefly discuss theoretical descriptions of networks (and their ensembles). We will continue by discussing various measures that have been applied to characterize the (structural) properties of networks before considering how the network affects properties of the nodes and vice versa, e.g. when analysing PIN data we may want to evaluate the extent to which the network shapes the evolutionary rate of the constituent proteins.

3. DESCRIBING THE STRUCTURE OF NETWORKS

We describe networks in terms of (static) graphs (Bollobás 1998); mathematically a graph is a pair of sets $\mathcal{G} = \{V, E\}$, where V is the set of N vertices or nodes and E the set of M (undirected) links or edges which connect pairs of nodes. Thus, each edge has an associated pair of vertices N_i and N_j (we will generally adopt the terminology used in the physics literature and also strive for a similar level of mathematical sophistication unless this may cause problems). Note that a node N_k may not have an associated edge, i.e. it may not be connected to any other node in the network; we also call

such nodes 'orphans'. A *connected component* is a set of nodes that is linked by edges but where no node in the component is connected to any node outside of the connected component. The largest component is often called the giant connected component.

Several representations for graphs exist but the conceptually easiest is the adjacency matrix, A_g (Bollobás 1998; Albert & Barabasi 2002). For N nodes the entries, a_{ij}, of this $N\times N$ matrix are simply the number of edges between nodes i and j. For undirected graphs A_g is symmetric, $a_{ij}=a_{ji}$; for so-called *simple* graphs a_{ij} is either 0 or 1 and $a_{ii}=0$, i.e. multiple edges and edges beginning and ending on the same node are not allowed. As far as PINs are concerned present data do not allow us to specify either a direction or a weight to an individual edge; proteins may, however, interact with themselves and, therefore, non-zero diagonal elements of the adjacency matrix are possible. If a network consists of several components then it will be possible to write the adjacency matrix in block-form. Rows in the adjacency matrix which correspond to orphaned nodes will contain only the value 0 (Valiente 2002).

4. NETWORK STATISTICS

A number of statistics have been defined which seek to summarize structural properties of networks. These have been applied to both theoretical and real network data. We will now discuss them in some detail and outline their behaviour for both classical and scale-free random graphs. The description of networks is complicated by the fact that unlike regular lattices there is no real connection between nodes and their spatial relationship; a node has no spatial position *per se* (Evans 2004). From a statistical perspective it is perhaps interesting to note that there exists, to our knowledge, no sufficient (in a formal statistical sense; see, for example, Cox & Hinkley 1974; Silvey 1975) statistic for networks.

4.1. The degree distribution

The degree k of a node is the number of edges attached to it and the degree distribution $n(k)$ is the number of nodes of degree k for all $k\geq 0$ (Albert & Barabasi 2002; Newman 2003a). It captures the diversity of local neighbourhoods in the network. In a regular lattice like an d-dimensional hypercube or Caley-tree all nodes or lattice points will have identical neighbourhoods and the degree is simply the coordination number. In the Erdös–Rényi random graph it is possible to show that the number of edges attached to a node is given by

$$n(k) = Np^k(1-p)^{N-k}\binom{N}{k} \approx \frac{(Np)^k\exp(-Np)}{k!}, \quad (4.1)$$

i.e. for $N\to\infty$ the degree distribution takes on the form of a Poisson distribution with parameter N_p, the average degree in the network. In the scale-free network, the degree distribution takes on a power-law (Barabasi & Albert 1999),

$$n(k) = Nk^{-\gamma}/\zeta(\gamma), \quad (4.2)$$

where $\zeta(\gamma)$ is Riemann's zeta function, $\zeta(x)=\sum_{i=1}^{\infty}1/x^i$ for $x>1$ (Abramowitz & Stegun 1974). The term scale-free follows from analogy to concepts from statistical physics (in particular the theory of second order phase transitions) and the fact that for a pure power-law degree distribution the ratio $n(\alpha k)/n(k)$ depends only on α but not on k; there is no natural scale to the network. In many cases power-law-like behaviour is confined to the tails of the degree distribution. In order to identify power-laws we normally require that they last over at least two to three orders of magnitude (Jensen 1998; Sornette 2003).

The degree distribution describes only one aspect of the data and graphs with hugely different architecture can exhibit similar degree distributions: a tree can have the same degree distribution as a highly reticulated graph (Bender & Canfield 1978; Dorogovtsev *et al.* 2000; Burda *et al.* 2001). Nevertheless, it is perhaps the most commonly considered network characteristic. In particular, scale-free behaviour of networks is generally inferred solely from the shape of the degree distribution.

4.2. The clustering coefficient

The clustering coefficient (Watts & Strogatz 1998; Newman 2003b) is a measure of the average local neighbourhoods in a graph/network. It is defined as the probability that two nodes j and k which are connected to node i are themselves connected. For a node i with degree k_i there are $k_i(k_i-1)/2$ potential links among its direct neighbours. If K_i denotes the links actually observed among i's neighbours then the clustering coefficient of node i is defined by

$$c_i = \frac{2K_i}{k_i(k_i-1)} \quad \text{for } k\geq 2; \quad (4.3)$$

for $k<2$ we define $c_i=0$. The clustering coefficient of the total network is then given by averaging over all nodes, $c=(1/N)\sum_{i=1}^{N}c_i$. While it has become customary to show degree distributions rather than just the average degree, regrettably the same is not universally true for clustering coefficients. This is despite the fact that the clustering coefficients often vary quite considerably across a network (Newman 2001). Studying such variation, for example, reveals that most nodes have a small clustering coefficient $c_i\approx 0$. This reflects the observation that local neighbourhoods of nodes in a network are often tree-like.

4.3. Average path-length and network diameter

As the position of nodes in networks bears no relationship to their spatial positions the distance between two nodes i and j is defined through the minimum number of edges that have to be traversed to reach j starting from node i. For directed networks the shortest path in the network may be the distance between i and j and j and i, respectively; in undirected networks it is of course the same. If we denote the distance between nodes i and j by l_{ij} then the average path-length is defined by

$$\langle l\rangle = \frac{2}{N(N-1)}\sum_{\langle i,j\rangle} l_{ij}, \quad (4.4)$$

where $\langle i, j \rangle$ indicates that the sum runs over all different pairs i, j (Valiente 2002).

The diameter of a network is given by the maximum distance in the network, i.e.

$$D = \max(l_{ij}). \tag{4.5}$$

If in an undirected network there is no path connecting nodes i and j then the distance l_{ij} is set to ∞. This is, for example, the case if the network consists of a number of connected components whence the average path length and the network diameter are also defined to be ∞. Unlike the previous statistics average path-length and network diameter are computationally quite intensive. Calculating all shortest paths in a graph is at least of order $O(N^2 \ln(N))$ (if we use adjacency lists).

These statistics have attracted considerable interest following the work of Watts, Strogratz (Watts & Strogatz 1998) and others (Newman & Watts 1999; Newman 2000; Zhou 2002) which reevaluates Stanley Milgram's classical notion of six-degrees of separation (Milgram 1967; Travers & Milgram 1969). Briefly, social networks, and most biological networks have a much smaller diameter than would be naively expected. In a regular one-dimensional network the diameter is given by the number of nodes N; in a ring the diameter is equal to $N/2$ and in a square lattice the largest distance is given by $2\sqrt{N}$. For real networks it was, however, observed that the diameter (or equivalently the average path-length) increases very slowly, e.g. $D \propto \log(N)$.

There appears to have been some misunderstanding (Bader *et al.* 2004) about just how common short average pathlengths or small network diameters are in real networks as well as in theoretical network models: classical random graphs (above the structural phase-transition) have this property as do (many) scale-free networks and virtually all naturally occurring networks. Thus, small world effects are the rule, not the exception (Watts 1999; Newman 2000). In a strict sense small-world behaviour (e.g. as exemplified by the models of Watts 1999; Newman 2000) requires a logarithmically growing diameter (or average path-length) and a high clustering coefficient.

4.4. Network motifs

Alon and co-workers (Milo *et al.* 2002, 2004; Shen-Orr *et al.* 2002; Kashtan *et al.* 2004) have introduced the notion of network motifs. In their definition (and the term has been applied differently by other authors) a motif is a pattern that occurs at a statistically increased frequency in the network. In part *a* of figure 2 we show the possible motifs that can occur between three nodes in a directed network; part *b* of the same figure shows the four-node motifs in an undirected network.

In the search for motifs one first counts all occurrences of the various patterns of interest. The statistical significance of each pattern is then assessed by randomizing the edges in the true network among the nodes, keeping the degree of each node fixed to its observed connectivity; the frequencies of the patterns

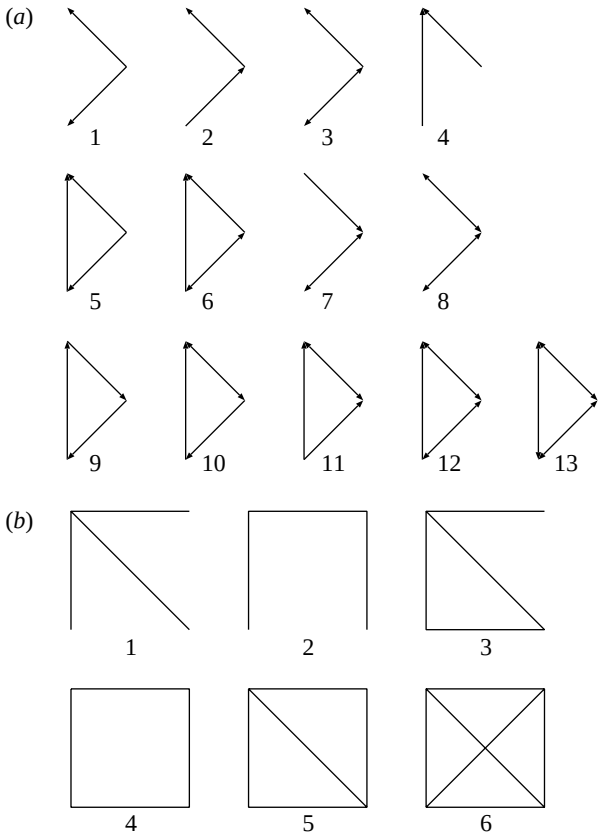

Figure 2. (*a*) all possible motifs defined by three nodes in a directed network. (*b*) All motifs possible defined by four nodes in an undirected network.

are then determined for the randomized network. Repeating this for a sufficiently large number of times yields a frequency distribution for each pattern in the ensemble of randomized networks. From this it is possible to arrive at a p-value for the pattern in the true network. Those patterns that occur at a significantly increased frequency in the true network are called motifs (Milo *et al.* 2002; Maslov *et al.* 2003).

It is worthwhile to consider what the meaning of these motifs is; this is, in fact, somewhat easier to see in directed networks where a direct and intuitive analogy to logic or electronic circuits exists. For example, the pattern 9 shown in part *a* of figure 2 corresponds to a *feed-forward* loop. Scanning through a network may thus elucidate the regulatory architecture of the network. Alon *et al.* (Milo *et al.* 2004) have applied such an approach to study motif spectra of different networks and suggested that it is possible detect superfamilies of networks with similar motif-spectra, i.e. a similar local logical organization of the network. For undirected graphs, however, motifs will only tell us about the extent to which certain neighbourhoods are overrepresented in the network. In the case of PIN we may for example be able to determine how often quadruplets of proteins occur where each pair is interacting, pattern 6 in figure 2*b*.

While the notion of motifs is appealing, the interpretation of motifs and their biological relevance can be subject to some controversy. For example, Wuchty & Stadler (2003) find that proteins in highly

connected motifs are more evolutionary conserved than would be expected by chance.[1] Mazurie *et al.* (2005) on the other hand find no correlation between motifs and evolutionary or functional characteristics. These authors conclude that motifs cannot be analysed in isolation from the rest of the network.

4.5. Network spectra

The final, most detailed but perhaps hardest to interpret perspective on networks is provided by a network's spectrum (Chung 1997; Farkas *et al.* 2001; Albert & Barabasi 2002; Bu *et al.* 2003). This follows from the eigenvalues λ of the adjacency matrix $\boldsymbol{A}$, i.e. the solutions of $(\boldsymbol{A}-\lambda\boldsymbol{I})=0$, where $\boldsymbol{I}$ is the identity matrix. For a $N\times N$ adjacency matrix we will have N eigenvalues λ_i, $i=1, 2, \ldots, N$ and the spectrum of the adjacency matrix is defined by

$$\rho(\lambda) = \frac{1}{N}\sum_{j=1}^{N}\delta(\lambda-\lambda_j), \tag{4.6}$$

with $\delta(x)$ the standard Dirac delta function.

There has been considerable interest in the extent to which the spectrum of a graph reflects local and global network structure (Bu *et al.* 2003; Chen & Xu 2003; Kamp & Christensen 2003) and this will probably continue to be an area of interest.

5. THEORETICAL NETWORK ENSEMBLES

The analysis of real networks is greatly aided by understanding how the various statistics discussed in the previous section behave in theoretical models of networks (Krzywicki 2001). We will briefly outline the behaviour of, what have become, the two canonical ensembles of networks, the classical or Erdös–Rényi (Erdös & Rényi 1959, 1960) random graphs and the scale-free random graphs (Aiello *et al.* 2001; Bollobás & Riordan 2003).

Graph ensembles play a central role in the theoretical analysis of networks. They are defined by the following.

(i) A set of graphs $\mathcal{G}=(\mathcal{V},\mathcal{E})$.
(ii) A statistical weight $\wp(g)$ for each graph $g\in\mathcal{G}$.

If $\wp(g)$ is constant then the graph ensemble will be equivalent to the microcannonical ensemble of statistical physics. Similarly, for varying $\wp(G)$, network ensembles corresponding to the canonical (N and M fixed) and grand canonical ensembles (N fixed $M\geq 0$) can be constructed (Dorogovtsev & Mendes 2003).

5.1. Classical random graphs

There are two definitions of a classical random graph; these become identical in the thermodynamic limit, $N\to\infty$. The first, due to Erdös–Rényi (Erdös & Rényi 1959, 1960)and denoted by $G(N, M)$, is given by a set of N nodes and M edges which are randomly placed among the nodes; one may explicitly specify that there can be at most one edge between every pair of nodes but this is negligible until $M\simeq N(N-1)/2$. The second classical random graph ensemble was proposed by Gilbert (1959) and is here denoted by $G(N, p)$, where N is again the number of nodes and p is the probability of a pair of nodes being connected by an edge; in this ensemble the expected number of edges, $\mathbb{E}[M]=p\times N(N-1)/2$. The degree distribution of a classical random graph is given approximately a Poisson distribution (Binomial at small values of N) with parameter λ equal to the average number of edges per node.

Classical random graphs have been studied extensively in mathematics (Bollobás 1998; Janson *et al.* 1999) and statistical physics (Stauffer & Aharony 1992). Of particular interest has been the structural phase-transition in the thermodynamic limit $N\to\infty$. For $p\ll\tilde{p}=2/(N(N-1))$ the network or graph will consist of many separate small connected components. At $p=\tilde{p}$ one of these components grows, increasingly amalgamating with other smaller components; this is often referred to as the giant connected component. Quite generally classical random graphs exhibit the small-world property for $p>\tilde{p}$. This point has sometimes not been appreciated in parts of the literature, where network concepts have been applied to (especially) biological network data.

5.2. Scale-free random graphs

Many important real networks, including the molecular networks, have degree distributions which decay much more slowly than exponentially (Alm & Arkin 2003; Evangelisti & Wagner 2004; Li *et al.* 2004; Agrafioti *et al.* 2005). In terms of the degree distribution classical random graphs are therefore unable to explain at least some aspects of real data. Barabási and Albert (Barabasi *et al.* 1999*b*) have shown that a simple probabilistic model can give rise to networks with a fat-tailed degree distribution. The so-called Barabási–Albert model has since then been shown to be mathematically ill-defined by Bollobás & Riordan (2003), but the LCD construction[2], which gives rise to

[1]We note here that only the observed number of motifs is cited in Wuchty & Stadler (2003), not their Z-scores. Moreover in a network comprising 3183 proteins they find e.g. 3.6 million copies of motif 1 in figure 2*b*. This can only happen if motifs are counted in a highly degenerate way which raises the question as to whether such a motif definition will give rise to biologically meaningful results.

[2]In the LCD construction a graph is evolved as follows.
(i) Start from an empty graph $G_1^{(0)}$ at time $t=0$ which contains no nodes and no edges (we could also start from $G_1^{(1)}$, a network with a single node and a single edge which starts and ends at the same node).
(ii) Add a new node and attach it to node s with probability

$$P_s = \begin{cases} d_s(t-1)/(2t-1) & \text{for } 1\leq s\leq t-1, \\ 1/(2t-1) & \text{for } s=t. \end{cases} \tag{5.1}$$

where $d_s(t)$ is the connectivity of node s at time t.
(iii) Return to (ii).
The LCD construction allows (i) for nodes which are separated from the rest of the network, (ii) multiple edges between nodes and (iii) loops. Networks are grown by adding one node at a time. It is also possible to model processes where m edges are added at each time-step, $G_m^{(t)}$; $G_m^{(t)}$ is generated from $G_1^{(mt)}$by combining the first m nodes, $\nu'_1,\ldots,\nu'_m$ to form node ν_1, etc.

a properly defined graph ensemble can be used (Bollobás 1998). Essentially, scale-free random networks are generated through growing a network by adding a new node at each time step and attaching it to existing nodes proportional to their connectivity. Other growth models, in particular certain processes where existing nodes are duplicated, retaining their edges with probability $0<\pi<1$, can also give rise to scale-free models in the limit where $N\rightarrow\infty$ (Aiello *et al.* 2001). These networks are called scale-free because the ratio $\Pr(\alpha k)/\Pr(k)$ depends only on α but not on k. Scale-free networks capture some vestiges of real networks but the great attention they have received is also due to the fact that scale-free behaviour can be generated by relatively simple models (Barabasi *et al.* 1999*b*, 2001; Dorogovtsev *et al.* 2000; Goh *et al.* 2002; Moreira *et al.* 2002). They, too, are an oversimplification of the true process underlying network evolution (Yook *et al.* 2004).

5.3. Other random graph ensembles

Classical and scale-free random graphs have become the canonical theoretical exampled for real complex networks. Due to their shortcomings other models have been considered, too. These other models can be loosely divided into two classes: theoretically motivated models which generate networks that capture one or more aspect of real networks, e.g. an empirical degree distribution, or evolve networks by a more flexible mechanism (Dorogovtsev *et al.* 2002) and mechanistic models which implement a model of network growth or evolution.

The former approach was pioneered by Bender & Canfield (1978), and refined by Molloy & Reed (1995) and Newman and co-workers (Newman *et al.* 2001; Newman 2003*c*). The latter approach is more recent and examples include the various models that have been proposed to generate scale-free networks. Networks are generated by defining a set of nodes and the number of edges incident on them. These are then connected randomly to form a network with a predefined degree distribution.

More recently still, biologically motivated models have been introduced (Wagner 2003; Berg *et al.* 2004) where authors have looked at the basic biological processes involved in the generation of real biological networks. These may include:

Node duplication: existing nodes are duplicated and the copy retains some or all of the interactions of the original node. Levels of duplication can be estimated from phylogenetic comparisons of paralogues; initially, at least, duplicated genes/proteins would fulfil similar functions.

Node attachment: new nodes are added to the network and attached to existing nodes (preferentially or randomly); horizontal transfer may offer a corresponding biological mechanism.

Node deletion: existing nodes and their incident edges are deleted; this may for example happen if a gene incurs a loss-of-function mutation.

Edge dyanamics: new edges can be formed, existing edges deleted or rewired. Again, this may be caused by mutations to the coding sequence.

Based on these processes it is possible to derive properly defined network ensembles. These can either be parameterized by biological data or be used to estimate biological parameters such as the effective rate at which proteins where duplicated. It has to be kept in mind, though, that the true evolutionary process underlying networks was much more complicated and will have contained a number of unique events, e.g. the whole genome duplication event in *S. cerevisiae* about 200 million years ago. At the moment it is unclear if such contingent processes can be modelled by statistical network ensembles. As pointed our by Burda *et al.* (2001), however, even if a network ensemble does not capture the true dynamics of network evolution, the study of suitable network ensembles can provide insights into the probabilistic behaviour of networks (Berg & Lässig 2002; Burda & Krzywicki 2004). In light of the problems addressed above further studies into biologically motivated finite-size network models may offer more interesting insights into biological network than has been the case for scale-free networks.

6. ANALYSIS OF PROTEIN INTERACTION NETWORKS

We will illustrate some of the challenges posed by network data using protein interaction network data. The poor quality of such data has been well documented and there have been some attempts at improving data sets using *in-silico* methods or labourious curation. Ultimately, more reliable experimental techniques may offer the only option to arrive at more reliable data; however in evolutionary studies the mean of an observable is frequently overwhelmed by the corresponding variance. Thus, even given more reliable interaction data considerable statistical challenges will nevertheless persist and below we will outline three such areas.

6.1. Structural analysis

In figure 3 we show the degree distribution of the yeast PIN (circles) and the best-fit power-law model (red). The first observation is that present data differs quite substantially from the pure power-law. The deviation is most pronounced at low and high values of k. This is often believed to be due to finite-size effects (Dorogovtsev & Mendes 2003). Also shown in the figure is the best-fit lognormal distribution, which, interestingly, does a rather good job at describing the empirical distribution. The improved fit is confirmed using methods from formal statistical model selection theory. The curves in figure 3 were fitted by determining the parameters of two curves in a maximum likelihood framework. If we denote the log-likelihood (Davison 2003) of a (potentially vector-valued) parameter θ by $\mathrm{lk}(\theta)=\Sigma_k \ln(\Pr(k))$ then the Akaike information criterion (AIC) (Akaike 1983; Burnham & Anderson 1998) for model i is given by $\mathrm{AIC}_i=-2\mathrm{lk}(\theta_i)+2\nu_i$, where ν_i is the number of parameters of model i. Unlike the likelihood ratio test the AIC allows us to formally compare the expanatory power of non-nested probabilistic models. The AIC—or the related Bayesian

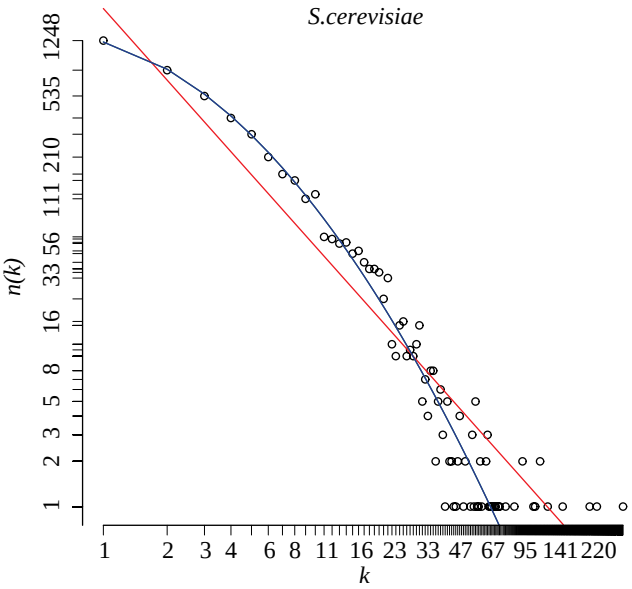

Figure 3. Degree distribution of the yeast protein interaction network (black circles) and best-fit power-law (red) and log–normal (blue) models (see Stumpf & Ingram 2005; Stumpf *et al.* 2005; Stumpf *et al.* in press for details).

information criterion—biases against models with more parameters that do not add significantly to a model's power to explain the data. While these methods frequently used in various branches of quantitative biology (see, for example, Strimmer & Rambaut 2002), they have not been widely applied in the analysis of network data.

The power-law has been favoured over other models, e.g. the Erdös–Rényi random graphs, because (i) it has a broader tail and (ii) simple mechanistic models asymptotically give rise to power-law degree distributions. Simulating the Barabási–Albert or LCD model we find that the AIC always favours the power–law distribution over the log–normal distribution (especially if analysis is restricted to connectivities $k \gtrsim 5$), even for small networks with $N = 500$; for real network data this does not appear to be the case. We find, that for the yeast PIN the lognormal distribution (blue) offers a better description of the data than the pure power-law or its heuristic finite-size versions (Stumpf & Ingram 2005; Stumpf *et al.* 2005, in press). The AIC for the power–law is $\approx 26\,920$ compared to $\approx 25\,430$ for the lognormal; when translated to relative likelihoods the lognormal is 10^5 more likely to explain the observed data than the power-law (Stumpf & Ingram 2005; Stumpf *et al.* 2005, in press). As all assertions of scale-free behaviour have been based on the degree distribution (Barabasi & Albert 1999; Wolf *et al.* 2002) this suggests that the notion of scale-free behaviour—as identified by a power-law degree distribution—may not hold up to statistical scrutiny. There have also been interesting attempts at using geometric random graphs (Penrose 2003) to describe networks; these also show that there are other network ensembles which may be better at explaining real biological networks than scale-free networks (Przulj *et al.* 2004).

The degree distribution captures only one, and by no means the most important, aspect of network data.

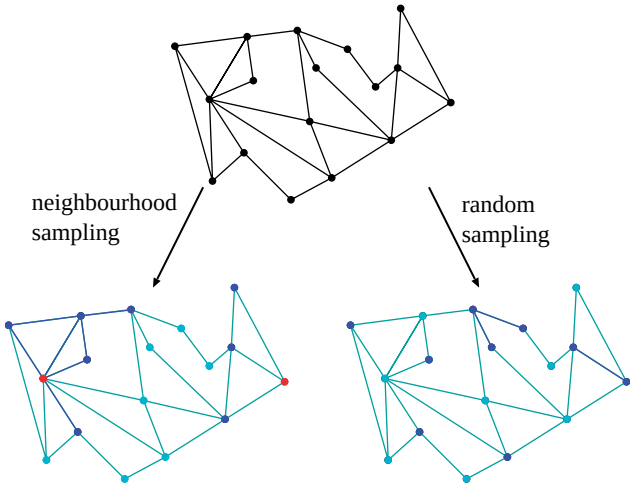

Figure 4. Two potential sampling schemes: under neighbourhood sampling, nodes connected to some initially chosen nodes (red) are more likely to be included in the set of nodes (blue) investigated. Under random sampling nodes are chosen (approximately) at random for the interaction studies. Lightblue nodes are not included in the experimental analysis.

Recently, Middendorf *et al.* (Middendorf *et al.* 2004, 2005) have developed approaches that aim to characterize networks formally and determine which ensemble best fits the data using machine learning techniques such as support vector machines. Their results confirm that there is more to networks than can be captured by scale-free ensembles; for example real networks tend to exhibit certain degree–degree correlations which standard network ensembles do not Berg & Lässig (2002)

6.2. Sampling properties

So far the vast majority of network analyses have considered the network data as potentially unreliable but, at least in principle, as representative of the true network. Until very recently (Stumpf *et al.* 2005) nobody had considered the fact that most network datasets, and certainly all biological network datasets, are only subnets embedded in the true but largely unobserved networks. The extent to which such subnets have identical or at least similar properties of the global network depends on the networks under different sampling schemes.

In figure 4 two different sampling schemes are illustrated. Under neighbourhood sampling, nodes are chosen (perhaps because of prior biological knowledge) and their local neighbourhood is explored; for example proteins involved in similar processes or in the same compartment may be included into the experimental set-up with higher probability. We would thus expect the local network neighbourhoods of the initial target nodes to fairly well represented. Under random sampling there is a finite probability $0 < p < 1$ for a protein to be included in the experiments. This is the most parsimonious sampling scheme as it does not require prior knowledge and/or bias of the experimenter. It also has by far the nicest mathematical properties and is most amenable to mathematical analysis (Stumpf *et al.* 2005; Stumpf & Wiuf in press).

For random sampling it is possible to determine whether the subnet will have similar properties as the overall global network; if this is the case we say the network is closed under the sampling scheme. For example, it is possible to show that the degree distribution of a subnet sampled from an Erdös–Rényi network will also be described by a Poisson distribution, but with parameter $p\lambda$ rather than λ. Interestingly, however, subnets from scale-free networks are not scale-free; in fact for most types of networks the subnet will be qualitatively different from the true network. The reverse is also true: if a subnet—for example, any of the current protein interaction network datasets—is shown to be scale-free, then the true global network cannot itself be scale-free. For neighbourhood sampling the situation is even more complicated: even Erdös–Rényi networks are no longer closed under this sampling scheme.

For networks of the type used by Berg *et al.* (2004) it can be shown, that they are also not closed under random sampling. This means inferences from network data, which do not take into account the fact that present molecular network data only describe subnets sampled from the whole network could be misleading. Graph spectra and inferences from motifs are even more susceptible to sampling error than the degree distribution. Interestingly, under random sampling the probability of retaining a motif in the subnet depends only on the number of nodes. Thus the probability of keeping any of the motifs in part *b* of figure 2 is p^4.

It is, however, relatively straightforward to adapt current network ensembles so that they incorporate sampling properties. If N' out of N nodes are included in the network data, then the ensembles can be used to model the evolution of a network of size N and then cull the nodes until a dataset of size N' is obtained. The network properties of subnets, thus constructed can then be compared to experimental data. Failure to appreciate the incomplete nature of networks, and the peculiar sampling features of networks, has compromised some published studies and some results may need to be reevaluated in light of the sampling process by which network data are collected.

6.3. Evolutionary and functional analysis

So far we have only discussed structural properties of networks. Here, we will assess the extent to which network structure reflects biological or biochemical processes or properties inside cells. In particular we are interested in the extent to which the network affects evolutionary properties of its constituent nodes (e.g. genes or proteins). Such system-level evolutionary information, in turn, will be informative about how transferable results of functional studies are between species.

An initial analysis of the impact of the interaction network on the evolutionary rate of proteins has been analysed by several groups (Pal *et al.* 2001; Wagner 2001; Fraser *et al.* 2002, 2003; Jordan *et al.* 2003; Fraser & Hirsh 2004; Bloom & Adami 2004) with different results. Some studies suggest that the evolutionary rate of proteins decreases with their connectivity (see, for example, Wagner 2001; Fraser *et al.* 2003) while others have suggested that this effect only applies to the most highly connected proteins, or becomes negligible once expression level differences have been accounted for (Jordan *et al.* 2003). These studies differed in the protein interaction data sets used, the species analysis and use of other data beyond interaction data. And all of these differences could have contributed to the different results obtained. A recent comparative study of the PINs in *S. cerevisiae* and *C. elegans* by Agrafioti *et al.* (2005), which used larger protein interaction data sets and larger, better resolved phylogenetic panels of closely species for rate estimation, as well as protein expression data and functional annotations reexamined this question. Expression appears to be indeed more closely correlated with changes in a protein's evolutionary rate than its connectivity. Present connectivity data does, in fact, have negligible explanatory power when other data such as gene ontology (GO) data is available. It has to be again kept in mind, though, that PIN data does not cover the whole protein interaction network but only the interactions between a subset of the proteins known to exist in *S. cerevisiae* and *C. elegans*.

Finally, there have been several studies which suggest that the properties, such as evolutionary rate of protein expression levels, of connected nodes are more similar than would be expected by chance (Williams & Hurst 2000; Fraser *et al.* 2002). The statistical significance of observed patterns is evaluated against a Null model of the network. Two different types of Null model have been used in the literature; the first has already been discussed in relation to network motifs (Maslov *et al.* 2003). In most evolutionary studies, however, a simpler bootstrapping (Efron & Tibshirani 1998) procedure was employed (e.g. see Fraser *et al.* 2002): if there are M edges in the original network then $2M$ nodes are chosen at random and their similarity calculated. Repeating this procedure a sufficiently large number of times yields an empirical Null distribution against which properties of the observed network are compared. This turn out to inflate the statistical significance of observed network data, as nodes with connectivity $k=1000$ contribute to the bootstrap samples with the same weight as nodes with connectivity $k=1$ or $k=0$. It has been shown that weighing nodes by their connectivity increases the 95% bootstrap confidence intervals considerably by up to 25% (Agrafioti *et al.* 2005).

More generally, the correct Null distribution for the statistical analysis of network data will depend not only on the structure of the network itself, but also its hierarchical organization (Yook *et al.* 2004): if proteins belonging to the same biological processes or located in the same cellular compartment are more likely to interact with proteins in the same process/compartment, then this ought to inform the Null model.

6.4. Comparing networks

Species comparisons have been used extensively to complement functional studies, including gene prediction and functional annotation of genes. A comparative approach, of network data from different species, for

Table 1. Number of orthologous pairs from reciprocal BLAST between organisms and number of shared interactions and motifs among shared orthologues.

(It appears that proteins in the same motif have related biological functions; for instance, all the proteins in the large yeast–human motifs appear to be related transcription initiation and transcription factors. PIN data was taken from the database of interacting proteins, dip.doe-mbi.ucla.edu. The small number of shared motifs reflects the incomplete nature of the interaction data as well as the difficulties in correctly identifying orthology.)

	orthologous proteins	shared edges	shared motifs
fly ↔ human	364	12	0
fly ↔ worm	1055	43	1×3-motif
worm ↔ human	228	4	0
yeast ↔ fly	1408	80	4×3-motif
yeast ↔ human	284	40	5×3-motif and 3×4-motif
yeast ↔ worm	692	23	1×3-motif

example like the one illustrated in figure 1, will provide insights into the functional organization at the system level and several groups have attempted to compare networks from different species. Most recently, Sharan *et al.* (2005) have identified conserved patterns of protein interactions in *S cerevisiae*, *C elegans* and *D. melanogaster* using sequence similarity to identify orthologous proteins. They have shown that such protein sequence-based network alignments can be used to assign protein functions and predict protein interactions. Unfortunately, such studies are plagued by the quality and the sampling nature of the data; therefore inferences are still quite limited. The numbers of orthologous proteins for which interaction data exists in more than one species is quite small (Kelley *et al.* 2003; Sharan *et al.* 2005; Wuchty & Almaas 2005). Using reciprocal BLAST searches to find putative orthologous proteins reveals that only a tiny number of patterns are shared between e.g. *D. melanogaster* and *C. elegans* where we find only a single pattern involving three nodes (see table 1). Recently, Sharan *et al.* (2005) have pointed out that straightforward reciprocal BLAST search does an insufficient job at reliably identifying shared patterns in different PIN; including additional potential orthologous proteins flagged up in the BLAST searches does, however, lead to the identification of pairs of proteins that (i) are sufficiently similar in sequence and (ii) appear to play a similar part in the PINs of different species. The identification of orthologous proteins between distantly related species—such as the species in table 1—will require the development of more powerful and rigorous inferential frameworks (Sonnhammer & Koonin 2002).

A complementary approach to such sequence-based approaches would be to align biological networks without reference to the sequence of the proteins/genes but based solely on the occurrence of certain patterns. The work by Berg & Lässig (2004) is a promising start in this direction; given the complexity of the graph isomorphism problem (Valiente 2002) and the lack of a well defined distance between graphs this is, however, a challenging problem. Combining and comparing resulting alignments will elucidate the impact of the network on the biology/biochemistry and vise versa. Biological information may also guide in the development of suitable heuristics for network alignment algorithms. One challenge here will be to construct an alignment procedure which integrates biological (sequence or GO) information, with strutural network properties.

7. CONCLUSION

Quantitative methods for the analysis of biological network data are still in their infancy. Given the scope and detail (however unreliable) simple statistical network models are reaching the limits of their applicability. Improved network models will have to be more flexible and take into account that networks are finite (and often too small for mean-field theories (Barabasi *et al.* 1999*a*; Newman *et al.* 2000) to be useful), have modular organization, and that present network data often contain only incomplete samples from the true network. Finally, they have to be more flexible at incorporating additional biological information. Unfortunately, this may entail using models with more parameters than the simple models used so far. Statistical model selection tools, like the AIC, will be useful in establishing sets of models which (i) can describe the data adequately and (ii) are not over-parameterized.

When describing biological processes at the system level (e.g. a cell) it is important to remember that the data is often very noisy, and the processes highly complex: molecular abundances and interactions change over time and in response to external stimuli as well as to dynamical intrasystem processes. The static language of graph theory used to describe todays biological networks may reach its limits when the conditionality and contingency of interactions need to be considered.

We thank Carsten Winf and Bob May for many helpful discussions; the manuscript has greatly benefited from the comments of three anonymous referees. Financial support from the Wellcome Trust is gratefully acknowledged.

REFERENCES

Abramowitz, M. & Stegun, I. 1974 *Handbook of mathematical functions*. New York: Dover.

Agrafioti, I., Swire, J., Abbott, I., Huntely, D., Butcher, S. & Stumpf, M. 2005 Comparative analysis of the *Saccaromyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. *BMC Evol. Biol.* **5**, 23. (doi:10.1186/1471-2148-5-23.)

Aiello, W., Chung, F. & Lu, L. 2001 A random graph model for power law graphs. *Exp. Math.* **10**, 53–66.

Akaike, H. 1983 Information measures and model selection. In *Proc. 44th Session of the Int. Statistical Institute*, pp. 277–291. Voorburg, The Netherlands: International Statistical Institute.

Albert, R. & Barabasi, A. 2002 Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97. (doi:10.1103/RevModPhys.74.47.)

Alm, E. & Arkin, A. P. 2003 Biological networks. *Curr. Opin. Struct. Biol.* **13**, 193–202. (doi:10.1016/S0959-440X(03)00031-9.)

Bader, J. S., Chaudhuri, A., Rothberg, J. M. & Chant, J. 2004 Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* **22**, 78–85. (doi:10.1038/nbt924.)

Barabasi, A. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509.)

Barabasi, A. & Albert, R. 1999*a* Mean-field theory for scale-free random networks. *Physica A* **272**, 173–187.

Barabasi, A., Albert, R. & Schiffer, P. 1999*b* The physics of sand castles: maximum angle of stability in wet and dry granular media. *Physica A Stat. Mech. Appl.* **266**, 366–371.

Barabasi, A., Ravasz, E. & Vicsek, T. 2001 Deterministic scale-free networks. *Physica A* **299**, 559–564.

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

Berg, J. & Lässig, M. 2002 Correlated random networks. *Phys. Rev. Lett.* **89**, 228 701. (doi:10.1103/PhysRevLett.89.228701.)

Berg, J. & Lässig, M. 2004 Local graph alignment and motif search in biological networks. *Proc. Natl Acad. Sci. USA* **101**, 14 689–14 694. (doi:10.1073/pnas.0305199101.)

Berg, J., Lassig, M. & Wagner, A. 2004 Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.* **5**, 51. (doi:10.1186/1471-2148-4-51.)

Bloom, J. & Adami, C. 2004 Evolutionary rate depends on number of protein–protein interactions independently of gene expression level: response. *BMC Evol. Biol.* **4**, 14. (http://www.biomedcentral.com/1471-2148/4/14.)

Bollobás, B. 1998 *Random graphs*. New York: Academic Press.

Bollobás, B. & Riordan, O. 2003 Mathematical results on scale-free graphs. In *Handbook of graphs and networks* (ed. S. Bornholdt & H. Schuster), pp. 1–34. New York: Wiley-VCH.

Bu, D. *et al.* 2003 Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res.* **31**, 2443–2450. (doi:10.1093/nar/gkg340.)

Burda, Z. & Krzywicki, A. 2004 Uncorrelated random networks. *Phys. Rev. E* **67**, 046118. (doi:10.1103/PhysRevE.67.046118.)

Burda, Z., Correia, J. D. & Krzywicki, A. 2001 Statistical ensemble of scale-free random graphs. *Phys. Rev. E* **64**, 046118. (doi:10.1103/PhysRevE.64.046118.)

Burnham, K. & Anderson, D. 1998 *Model selection and multimodel inference*. Berlin: Springer.

Chen, Y. 2003 Computational analyses of high-throughput protein–protein interaction data. *Curr. Protein Pept. Sci.* **4**, 159–181. (doi:10.2174/1389203033487225.)

Chung, F. 1997 *Spectral graph theory*. Regional Conference Series in Mathematics, vol. 91. Providence, RI: American Mathematical Society.

Cox, D. & Hinkley, D. 1974 *Theoretical statistics*. London: Chapman & Hall/CRC.

Davison, A. 2003 *Statistical models*. Cambridge: Cambridge University Press.

Dorogovtsev, S. & Mendes, J. 2003 *Evolution of networks*. Oxford: Oxford University Press.

Dorogovtsev, S., Mendes, J. & Samukhin, A. 2000 Structure of growing networks with preferential linking. *Phys. Rev. Lett.* **85**, 4633. (doi:10.1103/PhysRevLett.85.4633.)

Dorogovtsev, S., Mendes, J. & Samukhin, A. 2002 Multifractal properties of growing networks. *Europhys. Lett.* **57**, 334–338. (doi:10.1209/epl/i2002-00465-1.)

Efron, B. & Tibshirani, R. 1998 *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Erdös, P. & Rényi, A. 1959 On random graphs i. *Pubclicationes Mathematicae Debrecen* **5**, 290–297.

Erdös, P. & Rényi, A. 1960 On the evolution of random graphs. *Magyar Tud. Akad. Math. Kutató Int. Közl* **5**, 17–61.

Evangelisti, A. & Wagner, A. 2004 Molecular evolution in the yeast transcriptional regulation network. *J. Exp. Zool. B Mol. Dev. Evol.* **302B**, 392–411. (doi:10.1002/jez.b.20027.)

Evans, T. 2004 Complex networks. *Contemp. Phys.* **45**, 455–474. (doi:10.1080/00107510412331283531.)

Farkas, I., Derenyi, I., Barabasi, A. & Vicsek, T. 2001 Spectra of 'real-world' graphs: beyond the semicircle law. *Phys. Rev. E* **64**, 026704.

Felsenstein, J. 2003 *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.

Fraser, H. & Hirsh, A. 2004 Evolutionary rate depends on number of protein–protein interactions independently of gene expression level. *BMC Evol. Biol.* **4**, 13. (http://www.biomedcentral.com/1471-2148/4/13.)

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M. & Scharfe, C. 2002 Evolutionary rate in the protein interaction network. *Science* **296**, 750–752. (doi:10.1126/science.1068696.)

Fraser, H., Wall, D. & Hirsh, A. 2003 A simple dependence between protein evolution rate and the number of protein–protein interactions. *BMC Evol. Biol.* **3**, 11. (http://www.biomedcentral.com/1471-2148/3/11.)

Gilbert, E. 1959 Random graphs. *Ann. Math. Stat.* **30**, 1141–1144.

Goh, K.-I., Oh, E., Jeong, H., Kahng, B. & Kim, D. 2002 Classification of scale-free networks. *Proc. Natl Acad. Sci. USA* **99**, 12 583–12 588. (doi:10.1073/pnas.202301299.)

Janson, S., Lucaz, T. & Rucinski, A. 1999 *Random graphs*. New York: Wiley.

Jensen, H. F. 1998 *Self-organized criticality*. Cambridge: Cambridge University Press.

Jordan, I. K., Wolf, Y. I. & Koonin, E. V. 2003 No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* **3**, 1. (doi:10.1186/1471-2148-3-1.)

Kamp, C. & Christensen, K. 2003 Spectral analysis of protein–protein interactions in *Drosophila melanogaster*. Available at http://arxiv.org/abs/q-bio.MN/0405021.

Kashtan, N., Itzkovitz, S., Milo, R. & Alon, U. 2004 Topological generalizations of network motifs. *Phys. Rev. E* **70**, 031909. (doi:10.1103/PhysRevE.70.031909.)

Kelley, B., Sharan, R., Karp, R., Sittler, T., Root, D., Stockwell, B. & Ideker, T. 2003 Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA* **100**, 11 394–11 399. (doi:10.1073/pnas.1534710100.)

Krzywicki, A. 2001 Defining statistical ensembles of random graphs. Available at http://arxiv.org/abs/cond-mat/0110574.

Lappe, M. & Holm, L. 2004 Unraveling protein interaction networks with nearoptimal efficiency. *Nat. Biotechnol.* **22**, 98–103. (doi: 10.1038/nbt921.)

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.)

Li, W.-H. 1997 *Molecular evolution*. Sunderland, MA: Sinauer Associates.

Ma, H. & Zeng, P. 2003 Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277. (doi:10.1093/bioinformatics/19.2.270.)

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

Maslov, S., Sneppen, K. & Alon, U. 2003 *Correlation profiles and motifs in complex networks Handbook of graphs and networks*. New York: Wiley-VCH.

Mazurie, A., Bottani, S. & Vergassola, M. 2005 An evolutionary and functional assessment of regulatory network motifs. *Genome Biol.* **6**, R35. (doi:10.1186/gb-2005-6-4-r35.)

Middendorf, M., Ziv, E., Adams, C., Hom, J., Koytcheff, R., Levovitz, C., Woods, G., Chen, L. & Wiggins, C. 2004 Discriminative topological features reveal biological network mechanisms. *BMC Bioinform.* **5**, 181. (doi:10.1186/1471-2105-5-181.)

Middendorf, M., Etay, Z. & Wiggins, C. 2005 Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. *Proc. Natl Acad. Sci. USA* **102**, 3192–3197. (doi:10.1073/pnas.0409515102.)

Milgram, S. 1967 The small world problem. *Psychol. Today* **2**, 60–67.

Milo, R., Shen-Orr, 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.)

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.)

Molloy, M. & Reed, B. 1995 A critical point for random graphs with a given degree distribution. *Random Struct. Algor.* **6**, 161–179.

Moreira, A., Andrade, J. & Amaral, L. 2002 Extremum statistics in scale-free network models. *Phys. Rev. Lett.* **89**, 268703. (doi:10.1103/PhysRevLett.89.268703.)

Newman, M. 2000 Models of the small world. *J. Stat. Phys.* **101**, 819–841. (doi:10.1023/A:1026485807148.)

Newman, M. 2001 Clustering and preferential attachment in growing networks. *Phys. Rev. E* **64**, 025102.

Newman, M. 2003*a* The structure and function of complex networks. *SIAM Rev.* **45**, 167–256.

Newman, M. 2003*b* Properties of highly clustered networks. *Phys. Rev. E* **68**, 026121.

Newman, M. 2003*c* Random graphs as models of networks. In *Handbook of graphs and networks* (ed. S. Bornholdt & H. Schuster), pp. 35–68. New York: Wiley-VCH.

Newman, M. & Watts, D. 1999 Scaling and percolation in the small-world network model. *Phys. Rev. E* **60**, 7332–7342. (doi:10.1103/PhysRevE.60.7332.)

Newman, M., Moore, C. & Watts, D. 2000 Mean-field solution of the small-world network model. *Phys. Rev. Lett.* **84**, 3201–3204. (doi:10.1103/PhysRevLett.84.3201.)

Newman, M., Strogatz, S. & Watts, D. 2001 Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118. (doi:10.1103/PhysRevE.64.026118.)

Pal, C., Papp, B. & Hurst, L. 2001 Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931.

Penrose, M. 2003 *Random geometric graphs*. Oxford: Oxford University Press.

Przulj, N., Corneil, D. G. & Jurisica, I. 2004 Modeling interactome: scale-free or geometric? *Bioinformatics*. (doi10.1093/bioinformatics/bth436.)

Qin, H., Lu, H. H. S., Wu, W. B. & Li, W.-H. 2003 Evolution of the yeast protein interaction network. *Proc. Natl Acad. Sci. USA* **100**, 12 820–12 824. (doi:10.1073/pnas.2235584100.)

Ronen, M., Rosenberg, R., Shraiman, B. & Alon, U. 2002 Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl Acad. Sci. USA* **99**, 10 555–10 560. (doi:10.1073/pnas.152046799.)

Sharan, R., Suthram, S., Kelley, R., McCuine, S., Uetz, P., Sittler, T., Karp, R. & Ideker, T. 2005 Conserved patterns of protein interactions in multiple species. *Proc. Natl Acad. Sci. USA* **102**, 1974–1979. (doi:10.1073/pnas.0409522102.)

Shen-Orr, 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.)

Silvey, S. 1975 *Statistical inference*. London: Chapman & Hall.

Sonnhammer, E. & Koonin, E. 2002 Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **18**, 619–620. (doi:10.1016/S0168-9525(02)02793-2.)

Sornette, D. 2003 *Critical phenomena in natural sciences*. Berlin: Springer.

Stauffer, D. & Aharony, A. 1992 *Introduction to percolation theory*, 2nd edn. London: Taylor and Francis.

Strimmer, K. & Rambaut, A. 2002 Inferring confidence sets of possibly misspecified gene trees. *Proc. R. Soc. B* **269**, 127–142. (doi:10.1098/rspb.2001.1872.)

Stumpf, M. & Ingram, P. 2005 Probability models for degree distributions of protein interaction networks. *Europhys. Lett.* **71**, 152–158. (doi:10.1209/epl/i2004-10531-8.)

Stumpf, M. & Wiuf, C. In press. Sampling properties of random graphs: the degree distribution, *Phys. Rev. E*. (http://arxiv.org/abs/cond-mat/0507345.)

Stumpf, M., Wiuf, C. & May, R. 2005 Subnets of scale-free networks are not scalefree: the sampling properties of networks. *Proc. Natl Acad. Sci. USA* **102**, 4221–4224. (doi:10.1073/pnas.0501179102.)

Stumpf, M., Ingram, P., Nouvel, I. & Wiuf, C. In press. Statistical model selection methods applied to biological networks. *Trans. Comput. Syst. Biol.* (http://arXiv.org/abs/q-bio/0506013.)

Tong, A. H. Y. 2004 Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813. (doi:10.1126/science.1091317.)

Travers, J. & Milgram, S. 1969 An experimental study of the small world problem. *Sociometry* **32**, 425–443.

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

Valiente, G. 2002 *Algorithms on trees and graphs*. Berlin: Springer.

Wagner, A. 2001 The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283–1292.

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.)

Watts, D. 1999 *Small worlds*. Princeton: Princeton University Press.

Watts, D. & Strogatz, S. 1998 Collective dynamics of small-world networks. *Nature* **393**, 440–442. (doi:10.1038/30918.)

Williams, E. & Hurst, L. 2000 The proteins of linked genes evolve at similar rates. *Nature* **407**, 900–903. (doi:10.1038/35038066.)

Wolf, Y., Karev, G. & Koonin, E. 2002 Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays* **24**, 105–109. (doi:10.1002/bies.10059.)

Wuchty, S. & Almaas, E. 2005 Evolutionary cores of domain co-occurrence networks. *BMC Evol. Biol.* **5**, 24. (doi:10.1186/1471-2148-5-24.)

Wuchty, S. & Stadler, P. F. 2003 Centers of complex networks. *J. Theor. Biol.* **223**, 45–53. (doi:10.1016/S0022-5193(03)00071-7.)

Yook, S.-H., Oltvai, Z. N. & Barabsi, A.-L. 2004 Functional and topological characterization of protein interaction networks. *Proteomics* **4**, 928–942. (doi:10.1002/pmic.200300636.)

Zhou, H. 2002 Scaling exponents and clustering coefficients of a growing random network. *Phys. Rev. E* **66**, 016125.