Review TRENDS in Ecology and Evolution Vol.22 No.7

Evolution at the system level: the natural history of protein interaction networks

Michael P.H. Stumpf[1,2], William P. Kelly[1], Thomas Thorne[1] and Carsten Wiuf[3]

[1] Division of Molecular Biosciences, Imperial College London, London, SW7 2AZ, UK
[2] Institute for Mathematical Sciences, Imperial College London, London, SW7 2AZ, UK
[3] Bioinformatics Research Centre, University of Aarhus, Aarhus Dk-8000 C, Denmark

Recent work leading to new insights into the molecular architecture underlying complex cellular phenotypes enables researchers to investigate evolutionary processes in unprecedented detail. Protein interaction network data, which are now available for an increasing number of species, promise new insights and there have been many recent studies investigating evolutionary aspects of these interaction networks, from mathematical studies of growing networks to detailed phylogenetic surveys of proteins in their interaction network context. Here, we review the spectrum of such approaches, and assess issues associated with analyzing such data from an evolutionary perspective. Currently, such analyses are statistically challenging, but could link present initiatives in systems biology with results and methodologies that have developed in evolutionary biology over the past 60 years.

Introduction

Over the past decade, networks have taken a prominent position in many different disciplines, from theoretical physics [1,2] and technology to sociology and the humanities [3]. In biology, they have gained particular prominence [4–6] and network-based descriptions now have a fundamental role in biology, particularly in systems biology (the attempt to combine system-wide biological information with predictive modelling). Their appeal might be because not only are they based on a rigorous mathematical framework (mostly graph theory [7] and statistical physics [8,9]), but they also provide a convenient graphical representation of complex processes (Figure 1). The role of network concepts in ecology, where networks have a longstanding history [10], and evolution has recently been reviewed [11,12]. Here, we focus on the evolutionary analysis of protein interaction networks (PIN; Box 1), a field of much activity over the past few years that also brings a distinctly evolutionary perspective to biological systems.

Progress in experimental systems biology provides us with data that enable interactions among molecules inside a cell to be measured. The collection, verification and validation of such data pose considerable statistical challenges, and together form an active field of bioinformatics research (e.g. Refs. [14–16]). Certainly, interactions will not occur all the time and under all conditions. Nevertheless, PIN data has attracted much attention because of the hope that understanding which proteins interact with one another will give us deeper insights into the molecular machinery underlying complex phenotypes.

It is best to consider a PIN as an averaged structure, which contains interactions that are realized at different times and/or under different conditions. In reality, however, our knowledge of the true PIN will be subject to uncertainty. Here, although we do not consider the collection of the data, we discuss how issues such as data quality can be addressed in the analysis of PIN data. From an evolutionary perspective, the data becoming available should enable researchers to explore the interplay between evolutionary (population-based) and molecular processes, particularly the extent to which molecular interactions affect evolutionary genetics.

Glossary

Centrality betweenness: the fraction of shortest paths in the network going through a node.
Classical random graph (Erdós-Renyi Random Graph): a random graph model with a binomial or poisson degree distribution.
Clustering coefficient: probability that two nodes *r* and *s*, which are neighbours to node I, are themselves neighbours.
Degree: number of interactions of a node or protein.
Degree distribution: function that specifies the frequency of nodes with degree k.
Diameter: the diameter of a network is the minimum distance among all pairs of nodes in the network.
Distance: in a network, the distance between two nodes *i* and *j* is defined as the minimum number of edges that have to be traversed to reach *j* from *i*. If there is no path between the nodes, then their distance is set to ∞.
Graph: mathematical representation of a network in terms of its set of nodes (proteins) and the set of edges (protein–protein interactions).
Motif: a set of nodes with a certain pattern of edges connecting them (e.g. a closed triangle among three nodes defines one motif of size 3).
Power law: functional relationship of the form $f(x) = cx^{-\gamma}$.
Random graph: mathematical object that specifies a probability for each given graph structure.
Scale-free network: a type of network that has a power law degree distribution.
Small-world effect: a network that has small average distance; some definitions also require that the clustering coefficient is much larger than that of an equivalent (same number of nodes and edges) classical random graph.
Subnet: a network that is part of a larger network. Subnets are obtained by choosing a subset of the nodes in the larger network and considering it together with the connections among this set of nodes.

Corresponding author: Stumpf, M.P.H. (m.stumpf@imperial.ac.uk).
Available online 1 May 2007.

www.sciencedirect.com 0169-5347/$ – see front matter © 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2007.04.004

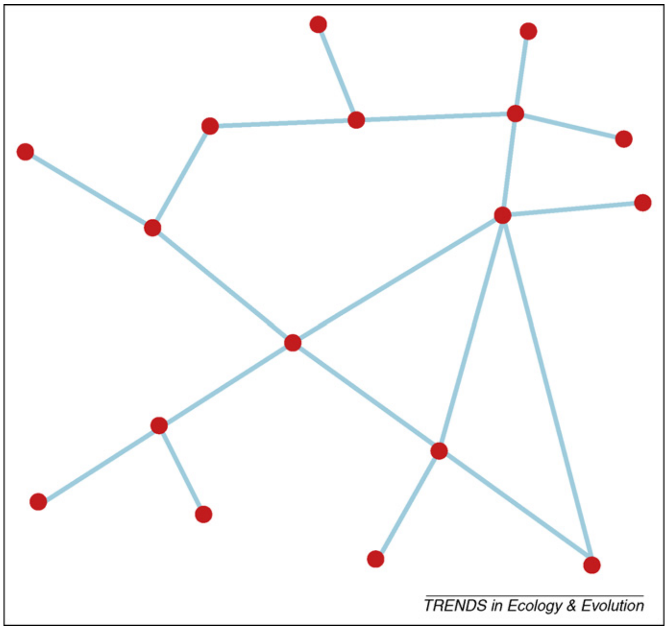

Figure 1. Example of a network and network statistics discussed in the main text. A network is generally described by a graph, *G*, which contains a set of nodes or vertices, *V* (red) and edges, *E* (cyan): thus, $G = (V,E)$. Here, we only consider undirected graphs with binary edges; that is, interaction between two proteins is either present or not; edges have no directions and no distinction is made between the relative strengths of different edges. In the future, quantitative interaction data will require straightforward extensions to the mathematical description of *G*. For recent reviews, see Refs [2,8,9].

The complexity of evolutionary analysis of biological networks is reflected by the diversity of different approaches used to study or model PIN evolution: from methods taken straight from statistical physics, via studies that involve methods from molecular evolution, to analyses that are heavily influenced by structural genomics. Here, we review these approaches as well as future challenges surrounding the evolutionary study of PINs.

From bags of genes to networks of interacting loci

The field of evolutionary genetics has made much progress in unravelling the molecular basis of genetic and phenotypic variation among individuals in a population, as well as among species. In particular, the interplay between theoretical analysis and experimental studies has led to the development of statistical frameworks for the quantitative analysis of genetic variation. At the level of populations of individuals belonging to the same species, population genetics and quantitative genetics have developed sets of extensively tested models for the evolution of systems consisting of a small and large number of genetic loci, respectively. These models have been studied carefully and, given a set of suitable assumptions, are amenable to exact mathematical analysis.

In population genetics, most studies focus on either a single locus or a few loci. Although for the former, our understanding of the model is now fairly complete [13,14], systems of interacting loci are an active field of interest, with many questions remaining. Most studies have looked either at pairs of loci or at systems of loci with certain simplifying limits, such as independent loci, where loci are in linkage equilibrium and are inherited independently. One crucial aspect of such theoretical models is the precise way in which the genotype is related to the phenotype (generally subsumed into some measure of darwinian fitness). The more that loci contribute to a trait, the more difficult modelling becomes, as additional assumptions have to be made: generally independence of the contributions from different loci is assumed. As the number of loci increases, however, systems enter the realm considered by quantitative genetics: here, a

Box 1. What are protein interaction networks?

Whereas metabolic networks and gene regulatory networks aim to summarize the basic biochemistry and the set of regulatory interactions of biological organisms, respectively, PINs lack such a straightforward interpretation. A PIN consists of all reported protein–protein interactions in an organism. When reporting an interaction between two proteins, we typically mean that some physicochemical interaction has been detected in *in vitro* biochemical assays, such as yeast-2 hybrid, immuno-precipitation and tandem-affinity purification, using protein tags. These experimental assays are subject to considerable noise levels, especially when used in high-throughput settings; thus, it is generally difficult to determine the extent to which interactions detected *in vitro* are relevant *in vivo*. Not all of these interactions will be realized simultaneously and there is as yet no data that would enable the analysis of protein interactions in the same organism under different environmental or physiological conditions. In general, the network data are also only of a qualitative nature; that is, interactions are either present or not but their strength is not quantified.

Finally, in reality, interactions are between different protein domains rather than proteins. Figure I shows the structure of the porcine pancreatic α-amylase (blue structure) in complex with a bean lectin-like inhibitor (red and yellow structure; protein database code 1DHK) [76]. The interaction occurs solely between the blue and red domains, although the inhibitor also has a 2nd domain, shown in yellow; other proteins containing the red and blue domains might also interact.

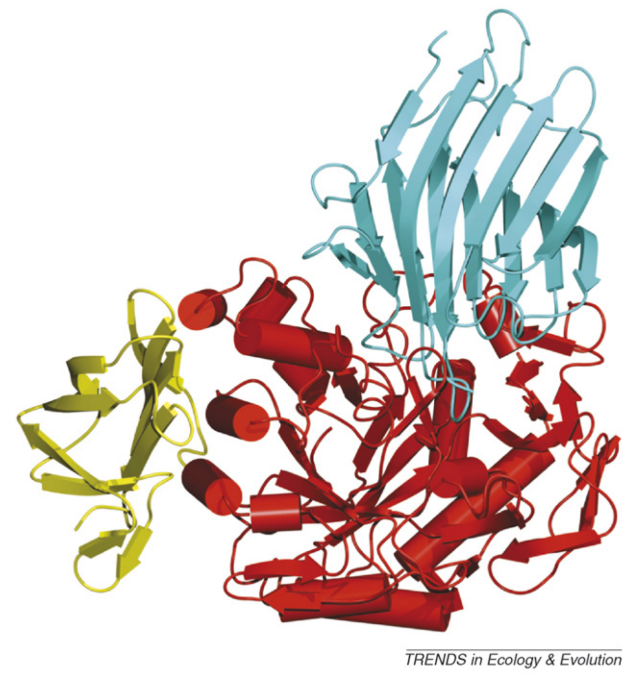

Figure I.

large number of genes are assumed to determine phenotypes, which, as a consequence vary continuously. The scenario of many (but not strictly infinitely many) interacting loci is largely uncharted territory and, to some extent, little progress has been made in this area since Haldane's classic paper 'In Defence of Beanbag Genetics' [15].

The evolutionary analysis of metabolic networks has perhaps the longest history, as metabolic networks are a natural extension of the biochemical pathways that have been studied for a long time. Following the pioneering work of Kacser and Burns [16,17], several studies have been published that combine models of biochemical pathways with explicit genetic models for the enzymatic activity or related phenotypes [18,19]. Here, the metabolic network is used as a map between genotype and phenotype to study the evolution of dominance and robustness.

As far as PINS are concerned, the available data are qualitatively different from what are typically modelled in population genetics: generally, a range of alleles at each locus and fitness schemes for collections of alleles are considered (e.g. Ref. [13]). In the analysis of PINs and metabolic networks, many studies refer solely to the viability of knockout mutants. The effects of mutations on the ability of proteins to interact and the resulting (epistatic) fitness effects have received relatively little attention.

Models of evolving networks

Much of the early theoretical work [20] used simple random graph structures as models of evolving networks. However, it became apparent that real networks have different properties. Notably, the degree of nodes in real biological, technological and social networks varies more than in classical random networks, with some nodes having very large degrees (so-called 'hubs') whereas most nodes have only a few interacting partners. This broad-tailed degree distribution cannot be achieved in classical random graphs (see Glossary), but several graph theoretical approaches have been developed that result in valid descriptions of such networks. Particularly interesting from the perspective of evolutionary and theoretical biologists are models of growing networks, which give rise to networks that have some of the desired properties of real

Box 2. Models of evolving networks

Models of growing random graphs aim to model the evolution of networks and three examples are illustrated in Figure I. Random growth by mechanisms such as attachment or duplication of nodes yields network structures that are different from those generated by an ensemble of networks [1] with the same number of nodes and edges but where the M edges are randomly distributed among the N nodes. This occurs even if the mechanism is subject to the constraint that the two resulting networks have the same degree distribution [77]. The process of growing networks by such processes introduces correlations among the nodes, and affects, for example, measures such as the clustering coefficient, which is higher than the expected clustering coefficient of a network with the same degree sequence but otherwise random connectivities.

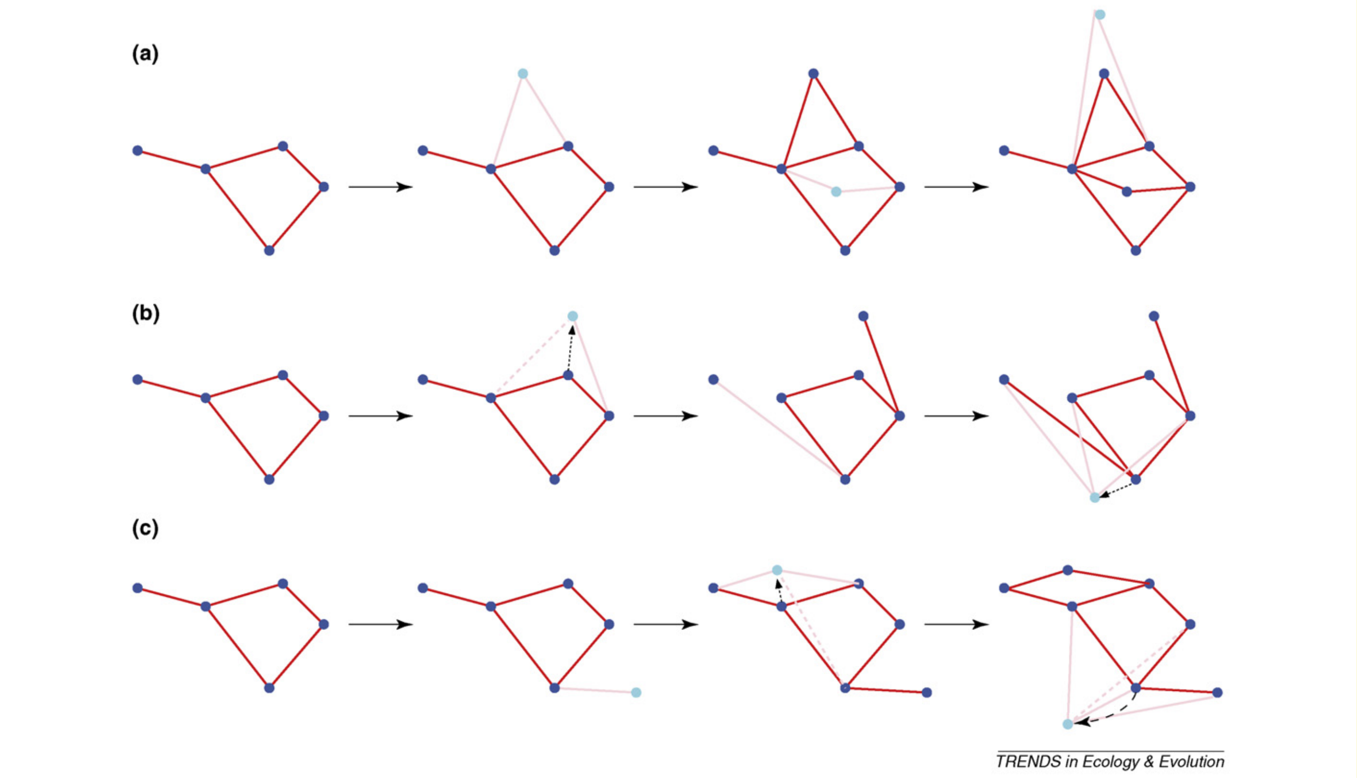

Figure I. Network growth models. Different mechanisms for network growth result in characteristic structures (new edges are shown in pink, new nodes in cyan). **(a)** corresponds to the preferential attachment model (at each time point, a new node is added and connected to two existing nodes, which are chosen in proportion to their degree); **(b)** illustrates the duplication divergence model (at each time point, either an existing node is duplicated and a random set of its interactions are inherited, or an interaction is rewired); **(c)** corresponds to the duplication attachment models (only duplication with random inheritance of interactions or preferential attachment events occur).

networks (e.g. broad degree distribution) and are based on models that might reflect or mimic some biological process of network evolution [21,22] (Box 2).

In the simplest case, such networks grow by adding a new node at each time-step and attaching it to randomly chosen nodes in the network. Such an approach will ultimately give rise to a degree distribution that decays exponentially with degree, k, for large degrees. The degree distribution of many real PINs, however, decays more slowly than exponentially. Attaching a new node to existing nodes with a probability proportional to their degree gives rise to networks with a much broader degree distribution. In fact, as the number of nodes n approaches infinity, the degree distribution of such networks will take on the shape of a power law, where the probability of a node having degree k is given by $\Pr(k) \propto k^{-3}$ [23].

The attachment process does not, however, correspond to a biological process (horizontal gene transfer is the only process that comes to mind, but is essentially limited to prokaryotes). Chung and co-workers [22] considered a model where, at each step, a node is chosen at random and duplicated. The new copy can either inherit all of the connections of the original, or it can inherit each edge with probability $0 < p < 1$. This, the authors argue, also gives rise to scale-free behaviour and a power law-like degree distribution with exponents $2 < \gamma < 3$, which would be in much better agreement with the observed power law exponents.

Statistically, however, the shape of the degree distribution is better described in terms of even simpler distributions (such as the log-normal) [24–28]. Some authors [29] have looked at network models that combine the processes of node duplication [22] with the potential for the properties of duplicated nodes to diverge from each other; these are sometimes referred to as duplication–divergence (DD) models. Related to this are models that consider network evolution by (preferential) attachment and duplication (with incomplete inheritance of edges) [30]. These attachment–duplication (AD) models and the DD models can also be analyzed in a quantitative manner, and fit the data equally well as the best phenomenological models (i.e. log-normal or stretched exponential distributions) and are based on a mechanistic model of network evolution that can be loosely (and perhaps simplistically) related to biological characteristics, such as rates of gene duplication. They do not, however, give rise to scale-free degree distributions [9,29] (Figure 2).

Although such models of network growth will offer oversimplified descriptions of the true process of network evolution, they can provide insights into general properties of evolving networks [4]. When combined with suitable statistical tools [27,30], it is possible to parameterize these models to obtain more realistic descriptions of evolving networks [6,31]. These models set out to achieve the same objective as models of sequence evolution [32,33].

Evolutionary analysis of networks

The relationship between the models mentioned above and actual PINs has been discussed in a largely qualitative fashion, although some approaches are informed by phylogenetic inferences [31]. With the wealth of available sequence data, however, it has become possible to study the effects of the network structure on the evolutionary properties of the constituent proteins. Before continuing, however, two notes of caution are needed, both of which pose immense statistical challenges to evolutionary analysis (even at the micro evolutionary level considered here):

(i) The evolution of proteins is known to be influenced by many factors; studying the effects of network structure on the rate of protein evolution without accounting for these potentially confounding factors could be misleading.

(ii) Present protein–protein interaction data are noisy and incomplete; failure to account for this might also introduce bias.

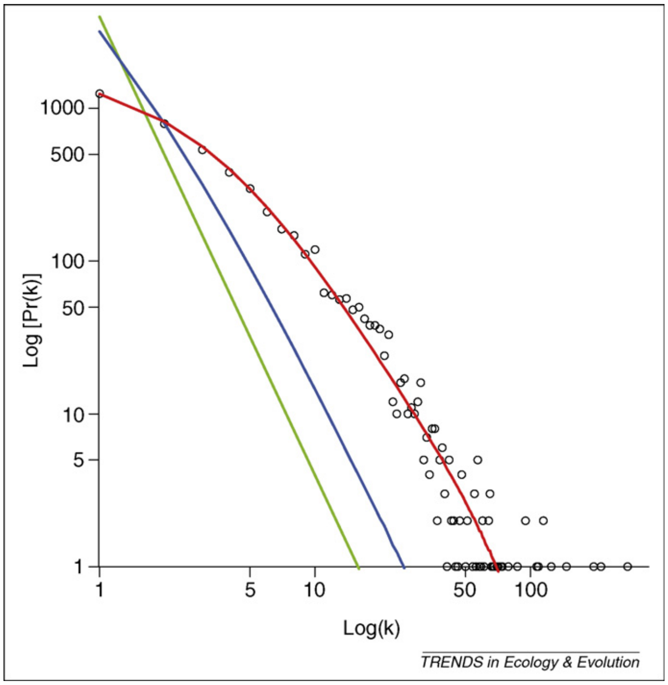

Figure 2. *Pr*(*k*) denotes the probabity of a node having *k* interactions. Network growth by the preferential attachment model (Box 2) up to the size of the observed network data results in a power-law degree distribution (blue); the degree distribution corresponding to the infinite network is shown in green. The attachment-duplication model (using maximum likelihood fits for the two free model parameters) results in the red curve; the observed degree distribution of *Saccharomyces cerevisiae* is given by the black circles. The model-based approach has the advantage that finite size is build into the model explicitly and that we can interpret the parameters biologically (duplication probability, π, and edge inheritance probability after duplication, *p*): the node duplication probability, for example, can be interpreted as a measure for the evidence for historical gene duplication, and can be compared with phylogenetic estimates.

Protein interaction data quality

The poor quality of high-throughput protein interaction data and the fact that the networks are incomplete (subnets generally have different properties from the true network) have been stressed repeatedly [34–39]. Simulation studies and comparative analysis of different data sets also show that the type of experimental approach underlying a data set influences the analysis [40]. There appears to be a genuine tradeoff between data quality and quantity that can be generated in medium–high throughput protein interaction assays. Nevertheless, many recent studies use only one data set from a single experiment or interaction database.

Given the uncertainty in the interaction data and the frequently weak statistical signal from evolutionary data, this is surprising and analyses might benefit from considering more than one data set; the new, extensively curated and validated data of Reguly *et al.* [41] should be a valuable addition. These authors collected 33 311 genetic and protein interactions from >31 000 print and online publications in yeast. These interactions were determined overwhelmingly in low-throughput and highly focused studies. We would therefore expect that the interactions are more reliable and less plagued by false-positive and false-negative results. Nevertheless, they still only capture a partial representation of the 'true' yeast PIN.

PINs and protein evolutionary rate

Several studies investigate the effect of network structure on the evolutionary rate of a protein [42–45], increasingly in connection with other factors such as expression level, biological annotations [e.g. those contained in the gene ontology (GO) resource, http://www.geneontology.org], lethality (of knockout mutants) and other factors [46–51]. Most of these studies have focused on the *Saccharomyces cerevisiae* PIN data, as it is the earliest and most comprehensive network. Comparing these studies is difficult because the measures for the evolutionary rate (e.g. inferred amino-acid substitution rates, numbers of synonymous and non-synonymous nucleotide changes or their ratio), the species used in the phylogenetic or comparative analysis, the PIN data sets and the statistical tools used differ (Table 1). The problem of making straightforward comparisons between different studies is further exacerbated by the fact that many of these characteristics are not independent and, similar to most evolutionary observations, show high variance.

All reported correlations between the evolutionary rate and protein degree have been relatively small (e.g. Fraser [42] found correlations of the order of 10%), and it has been claimed that, when correcting for expression level, this correlation is diminished further [47]. Recent and more extensive studies [49,50] confirmed this, and have found that measures for the expression level explain most of the variation in the evolutionary rate among proteins in yeast (as well as in *Caenorhabditis elegans* [49]). A recent study [50], which attempts to account for interdependency between different explanatory factors, demonstrates the predominance of gene expression measures over network statistics (e.g. degree and centrality). Perhaps the most reliable PIN data set available appears to exhibit no appreciable correlation between protein degree and evolutionary rate [41].

Interacting proteins, motifs and modules

The question of whether pairs of interacting proteins have more similar characteristics than do non-interacting proteins has also attracted interest [42,49,52–54]. Present data have the potential to open up a window on coevolution at the molecular level, but perhaps of more immediate importance is the possibility of using protein interaction data to annotate genes and their protein products (e.g. if a protein of unknown function interacts with one or more proteins of identical known function, it is a reasonable assumption that the 'new' protein will also have that function [55]). Quite rightly, such inferences are generally treated with caution, and are highlighted in the relevant ontologies to reflect the differences between inferences and (ideally) experimental knowledge.

From an evolutionary perspective, however, there is some evidence that properties of interacting proteins, such as their evolutionary rate [42], expression level [53] and regulatory elements [52], are more similar than would be expected by chance, although the correlation can often be weak [49]. The statistical significance of such results is assessed in the following way: (i) a measure of pairwise similarity (e.g. the correlation) of some property among

Table 1. Examples of studies that have analyzed the correlation between protein evolutionary rate and protein degree[a]

Evolutionary analysis	Network data	Results	Refs
Saccharomyces cerevisiae			
Evolutionary rate estimated between 309 (164 having interactions) 'well conserved' orthologues (350% identity) in *S. cerevisiae* and *C. elegans*	3541 interactions among 2445 proteins	r = -0.21, $P = 0.007$	[42,78,79]
dN rates from 1879 orthologous proteins between *S. cerevisiae* and *S. pombe*	1004 proteins from orthologous set (MIPS)	$r^2 = 0.0065$, $P = 0.009$ (1004 proteins), $r^2 = 0.0003$, $P = 0.70$ (465 proteins with 340% identity)	[47]
dN/dS calculated from four species from *Saccharomyces* genus	555 proteins in network data	r = -0.403, $P = 5 \times 10^{-23}$ (whole data set), r = -0.277, $P = 3 \times 10^{-11}$ (conditional on expression)	[80]
Average evolutionary rate among sets of orthologous proteins between *S. cerevisiae* and six other yeast species	15461 interactions among 4773 proteins (DIP)	Kendall's rank correlation: $\tau = -0.06$ (reduced when taking partial correlation corrected for expression level)	[49]
Caenorhabditis elegans			
Average evolutionary rate from pairs orthologous proteins between *C. elegans* and *C. briggsae*	7221 interactions among 2386 proteins (DIP)	$\tau = -0.03$, $P < 10^{-10}$ (reduced when taking partial correlation corrected for codon adaptation bias)	[49]
Drosophila melanogaster			
dN rates of 8748 gene pairs, obtained using PAML, between *D. melanogaster* and *D. pseudoobscura* and 1255 gene pairs between *D. melanogaster* and *Anopheles gambiae*	4625 high-confidence interactions	r = -0.10, $P = 0.01$	[83]

[a]Abbreviations: DIP, Database of Interacting Proteins [81]; dN, rate of non-synonymous substitutions; dN/dS, ratio of non-synonymous:synonymous substitutions; MIPS, Munich Information Centre for Protein Sequences [82]; r, Spearman Rank correlation coefficient; r^2, least square's regression coefficient; τ, Kendall's rank correlation coefficient.

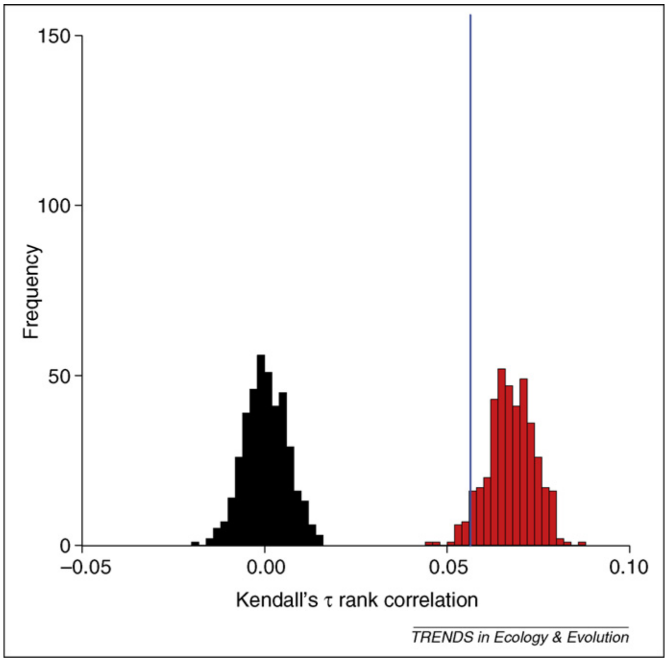

Figure 3. Correlation (measured by Kendall's τ rank correlation coefficient) of the evolutionary rates of interacting proteins (blue line) [49]. The two histograms represent the distributions of expected values for this correlation under null models of network organization, which condition only on the degree sequence of the network (black) and degree sequence and patterns of interactions among proteins with different GO annotations (red). Including the available annotations into the statistical null model results in a distribution that covers the experimentally observed average correlation among the evolutionary rates of interacting proteins.

pairs of interacting proteins is evaluated; (ii) the edges in the network are randomly rewired (the degree of each protein is kept fixed); and (iii) the statistic is calculated for the rewired network. Repeating the rewiring enough times results in a distribution that corresponds to a null model of the network, in which correlations among nodes have been removed. The distribution of expected correlations under this null model is thus symmetrically centred around zero (e.g. the black histogram in Figure 3) and this model can typically be rejected by the observed levels of correlation. Interestingly, however, when this rewiring approach is conditioned on other available data, such as the observed pattern of within and between GO, we end up with a different, improved (in the sense that they reflect the available biological information) null model. For yeast, the observed value lies within the distribution of the new null model. This observation reflects the hierarchical and modular organization that is inherent to biological networks [56–59]: averaged over the whole PIN, proteins are more likely to interact with proteins of a similar type than with proteins with different properties.

Modularity has attracted great interest in evolutionary and systems biology; it leads to networks that are no longer homogeneous and for which simple notions of network ensembles from theoretical physics might be inadequate. Studies of evolutionary properties in terms of network motifs, however, are complicated by the fact that motifs are generally defined statistically [60,61]. Therefore the same node might occur in many different instances of the same motif and simply counting the motif occurrences will result in enormous numbers [44] of some motif instances. Other definitions of motifs exist [62], which are more closely related to network communities [63] and their evolutionary behaviour can perhaps be analyzed more straightforwardly [64]. Certainly, for *Escherichia coli*, motifs in signalling networks appear to have evolved to balance performance with the cost of providing the molecular machinery necessary to guarantee accurate signalling [65].

A structural perspective on the evolution of PINs

The studies discussed so far have considered evolution in terms of amino acid or nucleotide sequences. There is, however, much evolutionary work from a structural (especially protein structure) perspective [66,67]. Strictly speaking, interactions are not between proteins but between protein domains [68]. Thus, it will be preferable to consider domains and structural properties directly. In particular, the detailed analysis of protein complexes [69] is aided by structural genomics and bioinformatics approaches; considering domains might also elucidate whether protein complexes are formed by pairwise interactions among their constituent proteins, or if they are due to collective interactions.

Caveats

In addition to the issues related to quality and completeness of protein interaction data already discussed, there are several other problems that need to be considered when analyzing PIN data in an evolutionary framework. First, generally in evolutionary biology, the variance of most estimators exceeds the mean and assessing the variation of an estimate is imperative. Second, many of the quantities that have been used to explain the variation in evolutionary rate among proteins are closely related [50] and the dependencies have to be accounted for; this is, for example, the case for different measures of protein abundance.

Many protein characteristics are also context dependent: essentiality of a gene depends on the environment and it is unclear to what extent knockout studies in laboratory strains are meaningful as these organisms live in artificial and stable environments. For expression levels, we also have to take into account the conditions under which they were collected: only those conditions that will be encountered in the wild will have contributed to the natural selection shaping the genome of an organism. In multicellular organisms, this problem is further exacerbated by expression-level changes during development and in different tissues.

Finally, the PIN structure also changes with time and in response to developmental and external stimuli: the network analyzed is therefore a highly averaged and idealized structure. Not all interactions will be present at all times and the strengths of interactions will also change over time. Current experimental methodologies do not enable us to quantify interaction strengths in a medium or high-throughput fashion.

Comparative analysis of PINs

Ultimately, researchers would like to be able to compare networks, analogously to sequence comparison [32].

Although the formal comparison of networks or graphs is fraught with computational challenges [70], successful approaches for the alignment and comparison of biological networks are likely to use strategies similar to those used in sequence analysis. There are two essential quantities to consider: the similarity of the nodes (generally in terms of primary sequence); and the similarity of subgraphs. Taken together, these will enable researchers to infer how the network changes over evolutionary timescales between species, which could help studies of the effects of selection and drift at the system level.

Several approaches have been developed that seek to align biological networks [70–74], which use different heuristics. In essence, this always involves a relative weighting between network information and sequence similarity. Inferences will thus usually differ among approaches but there is already some preliminary evidence that such approaches can give novel insights. For example, a recent analysis of the malarial parasite *Plasmodium falciparum* PIN data and comparison with *S. cerevisiae* has highlighted differences in the molecular organization of *P. falciparum* compared with yeast [75].

A direct comparison between networks of distantly related organisms is unlikely to uncover more than the most basic evolutionary conserved interactions [6,75] and it is worth keeping in mind lessons learned in comparative genomics: our understanding of features of the genome, such as exon–intron boundaries and promoter regions has increased with the rapidly growing availability of whole-genome sequences of closely related species. In light of this, it is probably necessary to map PINs systematically in species related more closely to the model organisms for which we already have extensive PIN data. For *S. cerevisiae*, for example, having PIN data for *Candida glabrata*, *Candida albicans* and/or fission yeast, *Schizosaccharomyces pombe*, would give us a better handle to understand the evolution of PINS in yeast.

Conclusion

There are several aspects of protein interaction data that we could not review here that are also related to evolutionary biology: for example, evolutionary arguments are used to predict protein interaction data in other species based on homology arguments. Similar arguments can also be used to validate protein interaction data [35] and these have gained prominence in bioinformatics. Here, we have focused on modelling and analyzing the evolution of PINs, starting from an abstract mathematical level and proceeding to discussing recent advances in the evolutionary analysis and comparison of PIN data.

Evolutionary biologists are dealing with ever increasing amounts of biological data to elucidate the processes that gave rise to the species observed, and which have shaped their phenotypes. Systems biology promises to produce even more data and will provide molecular descriptions of many cellular and molecular phenotypes that can be used to augment the sequence data that has been the main resource for evolutionary biologists over the past few years.

Although the analysis of these data are challenging because of the quality and the complexity of such system-level data sets, the benefits can be substantial. Presently, as discussed above and partially summarized in Table 1, however, simple answers are likely to be contentious. Pooling and comparing protein interaction information (preferably increasing the quality of the experimental data) of different sources enables us to get at least some glimpses into the evolutionary history of PINs.

In addition to issues surrounding quality and completeness of interaction and phenotypic data, there is also considerable need for improving models of network evolution. In population and evolutionary genetics, the arguments for models, however simplified, are well rehearsed and widely accepted: models force researchers to specify hypotheses explicitly and enable them to be tested against real data. As shown in Figure 2, good agreement between models and data can already be achieved for relatively parameter-sparse models (with two or three parameters) once the finite nature of the network has been accounted for and no extraneous asymptotic assumptions are being made. Such mathematical models can: (i) provide better understanding of the generic features of evolving network structures; (ii) model qualitatively real data [31]; and (iii) serve as evolutionary models in biologically motivated network alignment procedures [74] similar to models of sequence evolution used in sequence alignment procedures. Ultimately, these models and analyzes will also help us to bridge the gap between population and quantitative genetics.

Acknowledgements

We thank Suhail Islam for help with the illustrations. Financial help from the Wellcome Trust is gratefully acknowledged by M.P.H.S., T.T. and W.K. M.P.H.S. and C.W. acknowledge financial support from the Royal Society and the Carlsberg Foundation. We thank E. de Silva, M. Sternberg and J. Swire for helpful discussions.

References

1 Burda, Z. *et al.* (2001) Statistical ensemble of scale-free random graphs. *Phys. Rev. E* 64, 046118
2 Evans, T. (2004) Complex networks. *Contemp. Phys.* 45, 455–474
3 Scott, J. (2000) *Social Network Analysis: A Handbook,* Sage Publications
4 Wagner, A. (2003) How the global structure of protein interaction networks evolves. *Proc. Biol. Sci.* 270, 457–466
5 Alm, E. and Arkin, A.P. (2003) Biological networks. *Curr. Opin. Struct. Biol.* 13, 193–202
6 de Silva, E. and Stumpf, M. Complex networks and simple models in biology. *J. R. Soc. Interface* (in press)
7 Bollobás, B. (1998) *Random Graphs,* Academic Press
8 Newman, M. (2003) The structure and function of complex networks. *SIAM Rev.* 45, 167–256
9 Dorogovtsev, S. and Mendes, J. (2003) *Evolution of Networks,* Oxford University Press
10 Sugihara, G. (1984) Graph theory, homology and food webs. *Proc. Symp. Appli. Math.* 30, 83–101
11 Proulx, S.R. *et al.* (2005) Network thinking in ecology and evolution. *Trends Ecol. Evol.* 20, 345–353
12 May, R.M. (2006) Network structure and the biology of populations. *Trends Ecol. Evol.* 21, 394–399
13 Bürger, R. (2000) *The Mathematical Theory of Selection, Recombination, and Mutation,* John Wiley & Sons
14 Ewens, W. (2004) *Mathematical Population Genetics,* (2nd edn), Springer
15 Haldane, J.B. (1964) A defense of beanbag genetics. *Perspect. Biol. Med.* 19, 343–359
16 Kacser, H. and Burns, J. (1973) The control of flux. *Symp. Soc. Exp. Biol.* 27, 65–104
17 Kacser, H. and Burns, J. (1979) Molecular democracy: who shares the controls. *Biochem. Soc. Trans.* 7, 1149–1160

18 Keightley, P.D. (1989) Models of quantitative variation of flux in metabolic pathways. *Genetics* 121, 869–876

19 Frank, S.A. (1999) Population and quantitative genetics of regulatory networks. *J. Theor. Biol.* 197, 281–294

20 Kauffman, S. (1993) *The Origins of Order,* Oxford University Press

21 Dorogovtsev, S.N. *et al.* (2000) Structure of growing networks with preferential linking. *Phys. Rev. Lett.* 85, 4633–4636

22 Chung, F. *et al.* (2003) Duplication models for biological networks. *J. Comput. Biol.* 10, 677–687

23 Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science* 286, 509–512

24 Thomas, A. *et al.* (2003) On the structure of protein–protein interaction networks. *Biochem. Soc. Trans.* 31, 1491–1496

25 Stumpf, M. and Ingram, P. (2005) Probability models for degree distributions of protein interaction networks. *Europhys. Lett.* 71, 152–158

26 Stumpf, M. *et al.* (2005) Statistical model selection methods applied to biological networks. *Trans. Comput. Systems Biol.* 3, 65–72

27 Tanaka, R. *et al.* (2005) Some protein interaction data do not exhibit power law statistics. *FEBS Lett.* 579, 5140–5144

28 Khanin, R. and Wit, E. (2006) How scale-free are biological networks? *J. Comput. Biol.* 13, 810–818

29 Dorogovtsev, S.N. *et al.* (2002) Multifractal properties of growing networks. *Europhys. Lett.* 57, 334–338

30 Wiuf, C. *et al.* (2006) A likelihood approach to the analysis of network data. *Proc. Natl. Acad. Sci. U. S. A.* 103, 7566–7570

31 Berg, J. *et al.* (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.* 5, 51

32 Durbin, R. *et al.* (1998) *Biological Sequence Analysis,* Cambridge University Press

33 Koshi, J.M. and Goldstein, R.A. (1998) Models of natural mutations including site heterogeneity. *Proteins* 32, 289–295

34 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417, 399–403

35 Bader, J.S. *et al.* (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* 22, 78–85

36 Stumpf, M.P. *et al.* (2005) Subnets of scale-free networks are not scale-free: the sampling properties of networks. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4221–4224

37 Stumpf, M.P. and Wiuf, C. (2005) Sampling properties of random graphs: the degree distribution. *Phys. Rev. E* 72, 036118

38 Han, J.D. *et al.* (2005) Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.* 23, 839–844

39 de Silva, E. *et al.* (2006) The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol.* 4, 39

40 Hakes, L. *et al.* (2005) Effect of dataset selection on the topological interpretation of protein interaction networks. *BMC Genom.* 6, 131

41 Reguly, T. *et al.* (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* 5, 11

42 Fraser, H.B. *et al.* (2002) Evolutionary rate in the protein interaction network. *Science* 296, 750–752

43 Fraser, H.B. *et al.* (2003) A simple dependence between protein evolution rate and the number of protein–protein interactions. *BMC Evol. Biol.* 3, 11

44 Wuchty, S. *et al.* (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.* 35, 176–179

45 Wuchty, S. (2004) Evolution and topology in the yeast protein interaction network. *Genome Res.* 14, 1310–1314

46 Bloom, J. and Adami, C. (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. *BMC Evol. Biol.* 3, 21

47 Jordan, I.K. *et al.* (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

48 Kunin, V. *et al.* (2004) Functional evolution of the yeast protein interaction network. *Mol. Biol. Evol.* 21, 1171–1176

49 Agrafioti, I. *et al.* (2005) Comparative analysis of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* protein interaction networks. *BMC Evol. Biol.* 5, 23

50 Drummond, D.A. *et al.* (2006) A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* 23, 327–337

51 Salathé, M. *et al.* (2006) The effect of multifunctionality on the rate of evolution in yeast. *Mol. Biol. Evol.* 23, 721–722

52 Manke, T. *et al.* (2003) Correlating protein–DNA and protein–protein interaction networks. *J. Mol. Biol.* 333, 75–85

53 Lemos, B. *et al.* (2004) Regulatory evolution across the protein interaction network. *Nat. Genet.* 36, 1059–1060

54 Makino, T. and Gojobori, T. (2006) The evolutionary rate of a protein in influenced by features of the interacting partners. *Mol. Biol. Evol.* 23, 784–789

55 Deng, M. *et al.* (2004) Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics* 20, 895–902

56 Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1128–1133

57 Gagneur, J. *et al.* (2004) Modular decomposition of protein-protein interaction networks. *Genome Biol.* 5, R57

58 Han, J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88–93

59 Yook, S-H. *et al.* (2004) Functional and topological characterization of protein interaction networks. *Proteomics* 4, 928–942

60 Milo, R. *et al.* (2002) Network motifs: Simple building blocks of complex networks. *Science* 298, 824–827

61 Milo, R. *et al.* (2004) Superfamilies of evolved and designed networks. *Science* 303, 1538–1542

62 Kuramochi, M. and Karypis, G. (2004) An efficient algorithm for discovering frequent subgraphs. *IEEE Trans. Know. Disc. Eng.* 16, 1038–1051

63 Girvan, M. and Newman, M. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7821–7826

64 Ingram, P. *et al.* (2006) Network motifs: structure does not determine function. *BMC Genomics* 7, 108

65 Kollmann, M. *et al.* (2005) Design principles of a bacterial signalling network. *Nature* 438, 504–507

66 Chothia, C. *et al.* (2003) Evolution of the protein repertoire. *Science* 300, 1701–1703

67 Orengo, C.A. and Thornton, J.M. (2005) Protein families and their evolution-a structural perspective. *Annu. Rev. Biochem.* 74, 867–900

68 Deng, M. *et al.* (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* 12, 1540–1548

69 Aloy, P. *et al.* (2004) Structure-based assembly of protein complexes in yeast. *Science* 303, 2026–2029

70 Berg, J. and Lässig, M. (2004) Local graph alignment and motif search in biological networks. *Proc. Natl. Acad. Sci. U. S. A.* 101, 14689–14694

71 Kelley, B.P. *et al.* (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. U. S. A.* 100, 11394–11399

72 Sharan, R. *et al.* (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. U. S. A.* 102, 1974–1979

73 Koyutürk, M. *et al.* (2006) Pairwise alignment of protein interaction networks. *J. Comput. Biol.* 13, 182–199

74 Berg, J. and Lässig, M. (2006) Cross species analysis of biological networks by Bayesian alignment. *Proc. Natl. Acad. Sci. U. S. A.* 103, 10967–10972

75 Suthram, S. *et al.* (2005) The *Plasmodium* protein network diverges from those of other eukaryotes. *Nature* 438, 108–112

76 Gilles, C. *et al.* (1996) Crystallization and preliminary x-ray analysis of pig porcine pancreatic alpha-amylase in complex with a bean lectin-like inhibitor. *Acta Crystallogr. D* 52, 581–582

77 Callaway, D.S. *et al.* (2000) Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Lett.* 85, 5468–5471

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

79 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574

80 Fraser, H.B. and Hirsh, A.E. (2004) Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evol. Biol.* 4, 13

81 Xenarios, I. *et al.* (2000) Dip: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291

82 Mewes, H.W. *et al.* (2005) Mips: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* 34 (Suppl. 1), D169–D172

83 Giot, L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736