Evolution of biomolecular networks — lessons from metabolic and protein interactions

Takuji Yamada and Peer Bork

Darwin200

Abstract | Despite only becoming popular at the beginning of this decade, biomolecular networks are now frameworks that facilitate many discoveries in molecular biology. The nodes of these networks are usually proteins (specifically enzymes in metabolic networks), whereas the links (or edges) are their interactions with other molecules. These networks are made up of protein–protein interactions or enzyme–enzyme interactions through shared metabolites in the case of metabolic networks. Evolutionary analysis has revealed that changes in the nodes and links in protein–protein interaction and metabolic networks are subject to different selection pressures owing to distinct topological features. However, many evolutionary constraints can be uncovered only if temporal and spatial aspects are included in the network analysis.

Graph theory
The study of the properties of graphs. A graph is a mathematical structure used to model the pairwise relationships between objects. It is composed of a collection of nodes (vertices) and links (edges).

TAP–MS
A method used to detect physical protein–protein interactions by a series of affinity column purifications, followed by mass spectrometry for their identification.

European Molecular Biology Laboratory, D-69117 Heidelberg, Germany.
e-mails: takuji.yamada@embl.de; bork@embl.de
doi:10.1038/nrm2787

Since the end of the 1990s, there has been a flood of interaction data for proteins and other biomolecules (FIG. 1a). Network representations of these data have allowed the application of graph theory to the biological data[1]. Hence, the proteins (the molecular 'parts lists') can be seen as network nodes and their interactions as network links (also called edges). Whereas information about the proteins themselves started to pour into databases in the last decade, with each completed genome sequencing project, large-scale data sets on protein interactions are being generated only in this decade.

In 2000, two groups independently published large-scale analyses of yeast protein–protein interactions (PPIs) using yeast two-hybrid methods[2,3], resulting in large networks of yeast proteins and providing new data for the emerging field of protein network analysis. Since then, PPI networks have been generated using two-hybrid approaches in *Helicobacter pylori*[4], *Caenorhabditis elegans*[5], *Drosophila melanogaster*[6] and humans[7,8]. Large-scale tandem affinity purification coupled to mass spectrometry (TAP–MS) came as a complementary technology and was applied on a large scale to *Escherichia coli*[9,10] and yeast[11–14], and other techniques such as the protein fragment complementation assay have also been scaled up for interaction detection in *Saccharomyces cerevisiae*[15].

Before it was used to describe PPIs in general, the concept of networks was used to describe metabolic reactions. For several species, meticulously curated high-quality networks were used for metabolic control analysis[16], which was only much later performed computationally[17]. In metabolic networks, the interactions are between enzymes catalysing consecutive reactions (called reaction maps or enzyme-centric networks; FIG. 1a) or between metabolites (substrates and products; called compound-centric maps)[18–20]. The genome sequencing projects have provided a boost towards completing the metabolic networks in several species. Detailed metabolic networks in *E. coli*, yeast and humans have been constructed on a genome-wide scale[21–23] and are accessible through several resources, for example BioCyc[24], KEGG[25] and Reactome[26].

The amount of PPI and metabolic network data is constantly increasing (FIG. 1b), and the availability of these data in several species enables comparative analysis to further our functional and evolutionary understanding of molecular and cellular processes. However, a proper comparison of these networks between different species also requires an understanding of the completeness and accuracy of the data included in the network, which are very difficult to estimate. Although the data on nodes (proteins) seem reasonably complete and correct, as deduced from genome projects (limits such as improper gene prediction, copy number variation or alternative splicing aside), the links (interactions) are still a matter of debate.

Currently, the best studied PPI networks come from *S. cerevisiae* (for example, there are 4,975 proteins and 17,612 interactions in the Database of Interacting Proteins (DIP)[27]; see Supplementary information S1 (box)). These networks include interactions between proteins

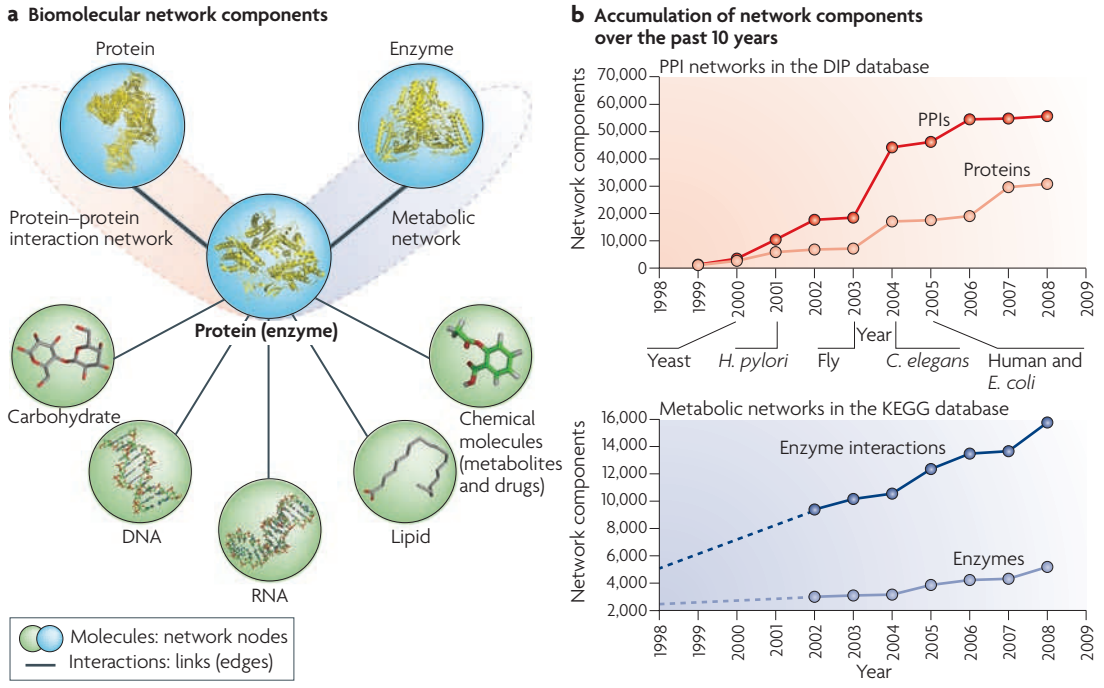


Figure 1 | **Components of biomolecular networks and their accumulation over time. a** | Basic components of molecular networks. Proteins are currently thought of as major nodes and their interactions are the links (also called edges). These protein–protein interactions (PPIs; orange ellipse) can be physical or indirectly derived — for example, by genetic means. The fraction of proteins that are enzymes and the interaction of these enzymes with other enzymes through shared metabolites represent metabolic networks (grey ellipse), which we compare here to PPI networks. Some proteins are involved in both PPI and metabolic networks (denoted by the merging of the two ellipses). The two network types discussed here have to be considered in the context of many other biomolecular network types. For example, regulatory networks deal with transcription factors and their interactions with DNA, and drug target networks refer to interactions between target proteins and chemicals. **b** | The number of parts lists (proteins) and links (physical interactions or enzymatic interactions) that have become available for different organisms during the past 10 years. Data on PPI and metabolic networks have been obtained from the DIP and KEGG databases[25,27,28,108].

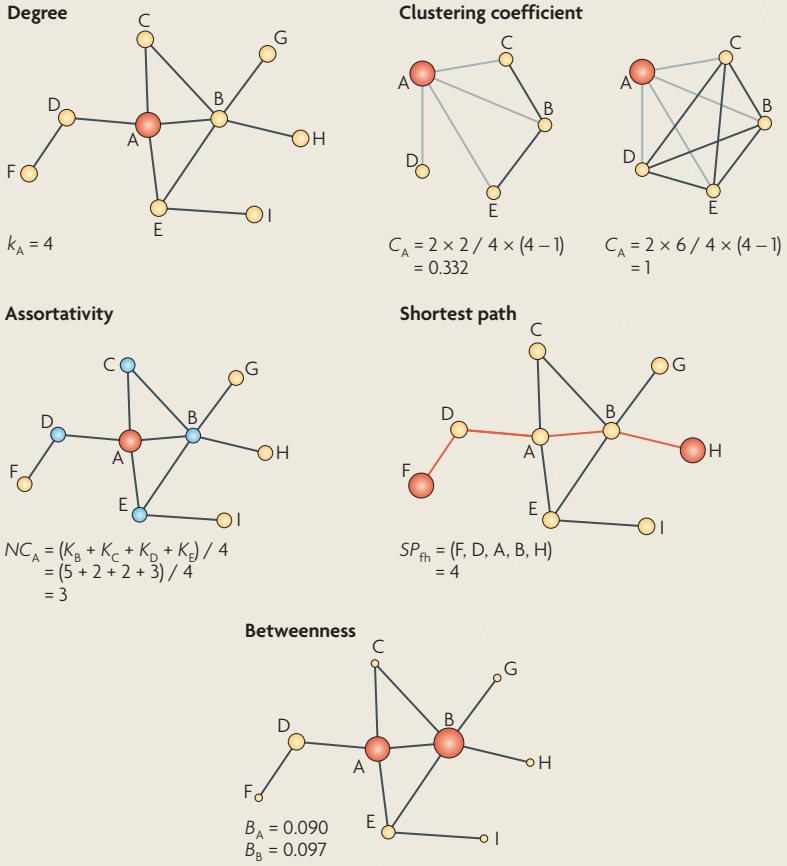
that function in stable protein complexes and more transient interactions that function, for example, in signalling pathways (some detection methods will not identify transient interactions). When integrating PPI networks with the currently available high-confidence data from various sources (see, for example, STRING, which is a database dedicated to protein–protein functional interactions[28]), the node coverage in yeast is above 80% (interaction partners have been reported for more than 5,000 yeast proteins), whereas the estimates on the total number of interactions (links) vary between 18,000 and 30,000 (REFS 29,30) (see Supplementary information S1 (box)). The number of links is therefore likely to be much larger than the current data suggest. The estimates of quality also vary, but it is noteworthy that a false positive rate for newly identified interactions of only a few percent was recently reported[15], which is considerably lower than in many other reports[31,32].

Although PPI and metabolic networks overlap to some extent, they can be distinguished by their distinct history (the purpose and use of the data and the way the networks were generated), how much is known about them and their accuracy. In contrast to PPI networks, metabolic networks do not come from a single large-scale project[33]. Instead, they are derived from numerous experiments carried out using different techniques and parameters in various laboratories. Thus, potential errors are of a fundamentally different nature and are less systematic than those of PPI networks (such as the under-representation of membrane proteins[29]). Curated metabolic networks are frequently taken as correct reference data sets, but they can also contain errors. For example, extensive experimental and computational studies of the small bacterium *Mycoplasma pneumoniae* (E. Yus, T. Maier, K. Michalodimitrakis, V. van Noort, T.Y., W.H. Chen, J.A.H. Wodke, M. Güell, S. Martínez, R. Bourgeois, S. Kühner, E. Raineri, I. Letunic, O.V. Kalinina, M. Rode, R. Herrmann, R. Gutiérrez-Gallego, R.B. Russell, A.C. Gavin, P.B and L. Serrano, unpublished observations) indicate a false positive rate of approximately 40% in databases such as KEGG (T.Y. and P.B., unpublished observations).

Despite different technical issues regarding data coverage and quality (see also Supplementary information S1 (box)) and despite the different nature of PPI and metabolic networks, we apply here comparative network analysis to discuss mechanisms of network evolution. For PPI we consider only physical interactions of proteins and for metabolic networks we include only curated data (as derived from the KEGG and Biocyc databases, which store manually curated metabolic networks and

Protein fragment complementation assay
A method used to measure protein–protein physical interactions. Protein interactions are coupled to the refolding of β-lactamase, which is fragmented and each of the two fragments is fused to the two proteins of interest. The reconstitution of β-lactamase activity acts as an interaction detector.

Box 1 | **Network measures**



Based on network components (nodes), local topologies of networks are characterized by several indices. In this box, we introduce the principal indices. Each network in the figure corresponds to red, yellow and blue nodes. Indices below each network are calculated for red nodes. Blue nodes represent the nearest neighbours of node A in the network for assortativity. Yellow nodes are the rest of the network components.

Degree (k). The number of neighbours of a node (also referred to as connectivity). For example, in the network shown in the figure, hypothetical node A has a degree of 4 ($k_A = 4$). The average degree of nodes for the whole network (<*k*>) is used as an index to describe the 'density' of a network. In networks in which each link has a selected direction, incoming (*k in*) and outgoing (*k out*) degrees need to be considered. The degree distribution ($p(k)$) gives the probability that a selected node has exactly *k* links.

Clustering coefficient (C). A measure of the degree of interconnectivity in the neighbourhood of a node[35]. In the network shown in the figure, the clustering coefficient for node A (C_A) is described as $2n_A / k_A(k_A - 1)$, where n_A is the number of links connecting the neighbours of node A to each other.

Assortativity (NC). The average degree of the nearest neighbours of a node[80] (see the figure). A negative correlation of assortativity with degree suggests that nodes that have a high connectivity (those in hubs, for example) tend to interact with nodes that have a relatively low connectivity. By contrast, a positive correlation suggests that the hubs tend to be located in highly connected topological modules.

Shortest path (SP). The path between two nodes in a network with a smaller number of steps than the many alternative paths between the two nodes. For example, in the figure, the shortest path from node F to node H (SP_{fh}) is composed of four steps.

Betweenness (B). A quantitative measure for describing the centrality of nodes in a network, provided as the frequency with which a node is located on the shortest path between all other nodes[78]. Nodes with high betweenness control the flow of information across a network. In the figure, the diameters of the nodes correlate with betweenness.

reactions). We do not consider a higher resolution of these networks. For example, some interactions can be attributed to particular amino acid residues and others to protein domains that form a more refined interaction network. To provide a basis for comparison, we first introduce general features of biological networks and describe elementary hypotheses for the evolution of PPI and metabolic networks. Then we discuss the evolution of network components (nodes and links) and their impact on network topology. Finally, we touch on attempts to go beyond two-dimensional networks by including spatial and temporal data that are becoming readily available.

General principles of network evolution

Networks have been studied for many years in different research fields, and the basic principles can be applied to biomolecular networks. In this section, we present basic features of and theories about biomolecular networks, with a focus on evolutionary aspects.

Topological descriptions of networks. A prerequisite for the study of network evolution is the quantification of their topological features. This can be carried out for the components, such as node degree (see below), and also for higher-order structures, such as scale-free structures. Topology is a concept from graph theory that is used to characterize the status of a network. Several measures are based on the components (BOX 1). For example, the degree (or connectivity; *k*) of a node — the number of links a node makes with other nodes — is the most elementary index (BOX 1). In networks, nodes with high degrees are called hubs. There are several other indices used to characterize network components (BOX 1). From these measurements the generic properties of a network can be derived; these, in turn, are the basis for distinguishing various network types (BOX 2).

Because it has been revealed that the topology of the metabolic network cannot be explained by random graph theory[34], many network types have been proposed to describe the global structure of biomolecular networks (BOX 2). Although many biological networks are described as being scale free (that is, their degree distribution approximates a power law), networks with hierarchical structures[35] that allow hubs and module structures within the network currently seem best suited for capturing most of the features of biological networks[36].

With these measures, the impact of the changes of network components can be quantified as they occur in molecular networks. Genetic changes can affect both the nodes (proteins) and the links (interactions) and networks can thereby evolve by addition or deletion of nodes and/or links. Taking this into account, several evolutionary models have been proposed for PPI and metabolic networks.

Network evolution by addition and loss of nodes and links. Current evolutionary models for network evolution are based on two types of genetic events. The first is essentially gene duplication and gene loss, which are probably the most important drivers of network evolution. Gene duplication implies the addition of a network node and

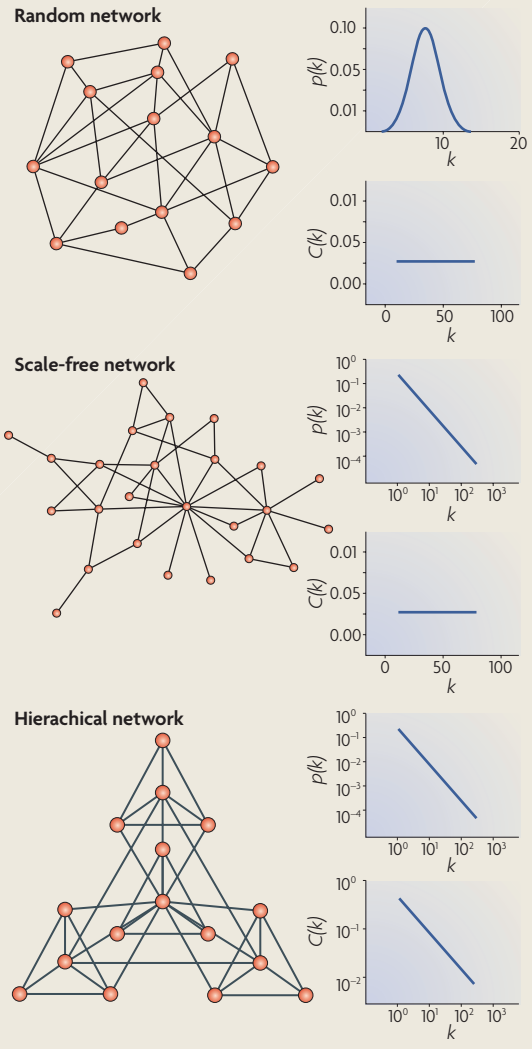
Box 2 | **Network topology types**

As well as local network measures based on components (see BOX 1), global structures of networks are also distinguished by several topological models. In this box, we introduce three representative models.

Random networks. Each pair of nodes is connected with probability *p*, which creates a network with randomly placed links[106] (see the figure). The node degrees follow a Poisson distribution, which indicates that most nodes have approximately the same number of links. The clustering coefficient (*C*) is independent of a node's degree (*k*), which means *C*(*k*) appears as a horizontal line if plotted as a function of *k*. In this network, the values for the shortest paths for most node pairs are relatively small, therefore a network with this feature is called a small-world network[36].

Scale-free networks. Networks that are characterized by a power law-like degree distribution[107] (see the figure). In a scale-free network, the probability that a node has *k* links follows $p(k) \sim k^{-\gamma}$, where γ is the degree exponent. Such distributions are seen as a straight line on a log–log plot. A relatively small number of highly connected nodes are known as hubs, and the probability of those hubs is statistically more significant than in a random network. In this model, the probability of an addional node connecting to an existing node depends on its degree[36].

Hierarchical networks. Networks with this structure allow hubs and modular structures to be inside the networks[35] (see the figure). To generate these networks, a densely linked starting module first has to be defined and replicated. The replicas of the module are then connected to the starting module. The replicas and starting modules are then used as the next starting modules, and so forth. The most important signature of hierarchical modularity is the scaling of the clustering coefficient, which follows $C(k) \sim k^{-1}$ and results in a straight line of slope −1 on a log–log plot. This structure seems to most closely reflect that of biological networks[35].



also the addition of links (BOX 3). With the loss of a gene, not only the node but also all associated links are lost. The second class of events are those that do not modify a gene as a whole, but modify a gene or its regulation in a way that results in link addition or loss. These genetic changes can be point mutations, insertions or deletions, or mutations that affect the regulation of the gene[37]. For example, non-synonymous nucleotide substitutions can easily modify the interaction interface of the encoded protein, leading to link addition or loss. The addition or deletion of a link then affects the connectivity of the network (BOX 3). As a result of a combination of these processes we can observe extensive network rewiring when comparing the interactions of homologous proteins[38] (BOX 3).

Metabolism-specific models for network evolution. As research in metabolism has been ongoing for many years, evolutionary models for metabolic networks have a much longer history than those for PPI networks[39–41]. Two models have been most influential: the retrograde model[42] and the patchwork model[20,43] (BOX 3).

In the retrograde model, pathways evolve backwards from a key metabolite. The principal concept of this model is that the environment is rich in metabolites that become initial key metabolites or intermediates. This assumption makes it difficult to explain the network evolution that occurs when organic molecules are depleted from the environment[42,44]. To overcome this limit of the retrograde model, the patchwork model was proposed. The patchwork model assumes an initial existence of broad specificity enzymes that become specialized after successive gene duplication events[20,43] (BOX 3).

Current data are more in support of the patchwork model. For example, enzymes containing triosephosphate isomerase (TIM)-barrel folds have been found in many different pathways, including carbohydrate metabolism, amino acid metabolism and nucleotide metabolism[45], and homologues are twice as likely to be found in different pathways than in the same pathway in *E. coli*[46]. Nevertheless, a high duplication rate of genes encoding enzymes catalysing consecutive reactions has been found[19,47], in concordance with the retrograde model. Taken together, these findings

Power law
A statistical model that describes that one quantity is proportional to the power of another quantity.

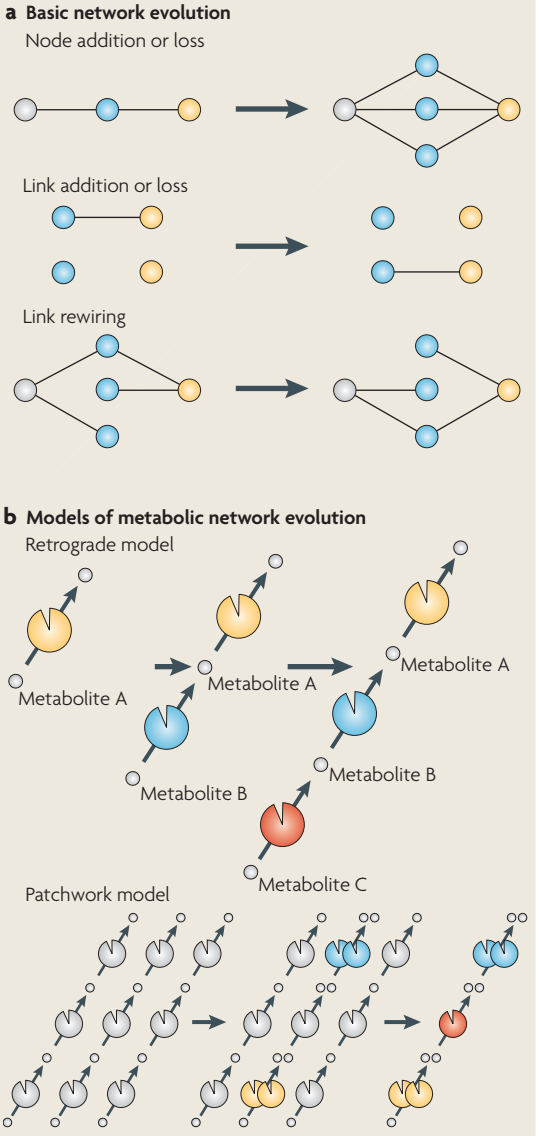
Triosephosphate isomerase (TIM)-barrel fold
The most frequent and conserved protein fold, comprising eight α-helices and eight β-sheets.

Box 3 | **Evolutionary models for PPI and metabolic networks**

The evolution of biomolecular networks is coupled to several genetic events. Node addition or loss in PPI and metabolic networks (see the figure, part **a**) usually implies that a gene duplication or loss has occurred and, implicitly, that the addition or loss of links has occurred, because each gene duplicate should keep all of its existing interactions. Horizontal transfer is another means of node addition, in which the impact on links can vary. Link addition or loss usually implies that genetic changes, such as point mutations, domain accretion or loss, alternative splicing, insertions or deletions, have occurred in genes or their regulatory regions. These genetic changes can destroy or create links. Link rewiring is usually a mixture of consequences from link addition or loss, often also involving secondary effects from node addition or loss. For the evolution of metabolic networks (see the figure, part **b**), environmental chemical conditions have to be considered. The following two models are the representative models that are specific to metabolism. The enzyme colours in the figure represent the order of recruitment: first yellow, second blue and third red.

Retrograde model. This model assumes that pathways are evolving backwards from a key metabolite. First, an organism that is heterotrophic for key metabolite A uses up all of the environmental supply of A. Second, the recruitment of an enzyme capable of synthesizing A from precursor B brings a selective advantage to the organism. In turn, environmental concentrations of B drop and this is compensated for by the recruitment of enzymes capable of synthesizing B from C.

Patchwork model. This model assumes that enzymes refine their substrate specificity after duplication. Initially, most of the enzymes have broad substrate specificities, which can catalyse multiple reactions. These broad substrate specificities of enzymes enable the generation of many metabolic pathways for the synthesis of the same key metabolites. The duplications of genes in such metabolic pathways bring selective advantage to the pathways because an increased level of the enzyme will generate more of the key metabolites. Finally, enzyme specialization following the gene duplication events will lead to the specialization of the different pathways.



imply that the models are not mutually exclusive but, instead, complementary. In any case, the selection pressure leading to network evolution is caused by external factors, in particular by environmental metabolites[48], which have to be considered together with population and community effects when discussing networks in organisms.

External factors in network evolution. Biological systems are not isolated, and therefore evolutionary processes must be affected by environmental conditions, such as nutrition, geochemical burdens, environmental stability and community effects. Current thinking in the scientific community is species-centric, but community effects that affect the influx and efflux of metabolites must influence mutational load and selection, which causes genetic changes. The metabolic network of an organism directly connects to the chemical universe of the external environment (which is often influenced by other organisms) and this must also have an impact on the evolution of the functional repertoire (the network parts lists)[49]. This is supported by the facts that essential sets of small molecules taken up from the environment differ among species[49] and that the functional repertoire of microbial communities as a whole adapts to varying geochemical conditions[50]. For example, the adaptation of metabolic networks to oxygen revealed the effect of this environmental variable on the architecture of the network: it seemed to be most effective to expand the enzymatic reactions that required oxygen[48].

Taken together, the network topology itself, as quantifiable by various measures, should be able to reveal at least some of the external selection pressures that cause network evolution, although currently relevant data are still sparse. Particular evolutionary constraints can be identified only when taking into account the background of the general principles of node and link evolution that we describe below.

Evolution of nodes (parts lists)

The evolution of network nodes (here proteins, specifically enzymes in metabolic networks) is coupled to the genetic material of a cell. In this section, we discuss the dynamics of network nodes, highlight some constraints on node evolution and elaborate on the impact of node addition or loss on network topology.

Evolutionary dynamics of nodes. Approximately 30% (1,312 out of 4,133) of genes in *E. coli* have a human orthologue and less than 5% (190) are conserved in more than 90% of completely sequenced species across the bacteria, archaea and eukaryota domains[29]. This reveals on the one hand that there is considerable conservation in organisms that have diverged more than four billion years ago, and on the other hand that node addition and loss are heavily contributing to network evolution and the core group of unchanged nodes is very small (that is, almost all nodes of the protein network are dispensable).

Whereas the loss of nodes (proteins) seems to be clock-like (it happens at a certain rate even in lineages in which genes are added to the genome of an organism), the rate of node addition varies a lot over time[51]. Several processes can contribute to the addition of nodes. The most frequent genetic event leading to node addition is gene duplication, which results from different genetic events such as whole-genome duplication, locally confined gene duplication and retrotransposition. Gene duplication probably influences PPI and metabolic network structures, with a major impact on organism phenotypes[38,52].

In organisms without a germ line (all organisms except higher eukaryotes), horizontal gene transfer can also contribute considerably to node addition and to the extent that entire genomes of organisms can be merged or acquired (as occured, for example, with mitochondria and chloroplasts) and the respective sub-networks be united. There are several other minor processes, such as domestication of phage or virus genes (for example, retrovirus genes becoming an integral part of the host genome through the deletion of flanking transposon sequences). Domestication of virus genes has contributed up to 1% of metazoan genes[53].

Constraints of node addition and loss. Although there are various node-specific constraints on gene loss (for example, essentiality) or gene duplication (for example, dosage effects of tightly controlled protein complex subunits that would impair the balance of links between nodes[54,55]), there seem to be network-associated laws with respect to the functional composition of genes in a network.

Depending on the network size (approximated by the number of genes in the genome of an organism), the fraction of genes from different functional categories is different[56,57], implying that there are constraints on the duplicability of genes. Genes in metabolism occupy a roughly constant fraction of a genome. However, many other categories show significant deviations. Whereas many genes associated with translation remain as a single copy regardless of the genome size, transcription factors, for example, increase exponentially with genome size such that their fraction is higher in larger genomes. These principles have implications for regulatory design in bacteria, but could also be used to predict the average genome size of samples from incomplete metagenomic data sets[58]. Despite selection against some kinds of node addition, most genes can be duplicated or horizontally transferred[59]. However, the effects can be different for rare whole-genome duplications or for horizontal transfers of many genes at once, owing to different topological structures of the networks (BOX 2; see below). The addition of individual nodes has different topological implications for PPI and metabolic networks.

Effects of node addition on PPI networks. Using a yeast PPI network, it has been shown that there is a negative correlation between protein connectivity and duplicability; that is, there is a direct relationship between the duplication rate of a gene (evolution of parts lists) and the network context of its encoded protein[60]. Highly connected proteins (hubs) therefore have a low duplicability. The stability of network structure mainly relies on these hubs; therefore, lethality increases threefold if a hub is deleted[61,62]. This is consistent with a slow evolutionary rate, implying that orthologues of hub proteins are highly conserved[63–65]. The duplication of a hub is therefore often deleterious because it affects a large number of proteins. However, although intuitive, this conclusion is not undisputed as some studies do not find any correlation between connectivity and duplicability[66] and others even find a positive one[54]. Measurements of evolutionary rates of proteins are also still controversial and the effects of protein connectivity seem to depend on the PPI network studied as well as on data coverage and quality. These topics therefore need further investigation.

By contrast, there is agreement that proteins interacting with the environment have a higher average duplicability than those that are localized within intracellular compartments. For example, the duplicability of proteins localized to the cell periphery, such as transporters, is higher than that of those localized in intracellular compartments, but their connectivity is lower. Conversely, proteins localized in the nucleus, such as enzymes for RNA and DNA metabolism, have a high connectivity and a low duplicability[60].

Effects of node addition on metabolic networks. As for PPI networks, the duplication rate of hubs in metabolic networks is relatively low[67], underpinning their central role. Not unexpectedly, duplication of genes (parts lists) seems to have mostly local effects in the network. Indeed, an over-representation of homologous enzymes in consecutive reactions has been shown and can be explained by a preferential biochemical coupling of these reactions. In addition, a high retention of duplicates between chemically similar reactions and in closely connected functional modules was observed[19]. The local connectivity effect of duplications could not be found in interaction networks of non-enzymatic proteins or gene transcriptional regulatory networks, suggesting that the retention of duplicates results from biochemical rules that govern substrate–enzyme–product relationships[19].

Orthologue
A gene present in different species that evolved from a common ancestral gene by speciation.

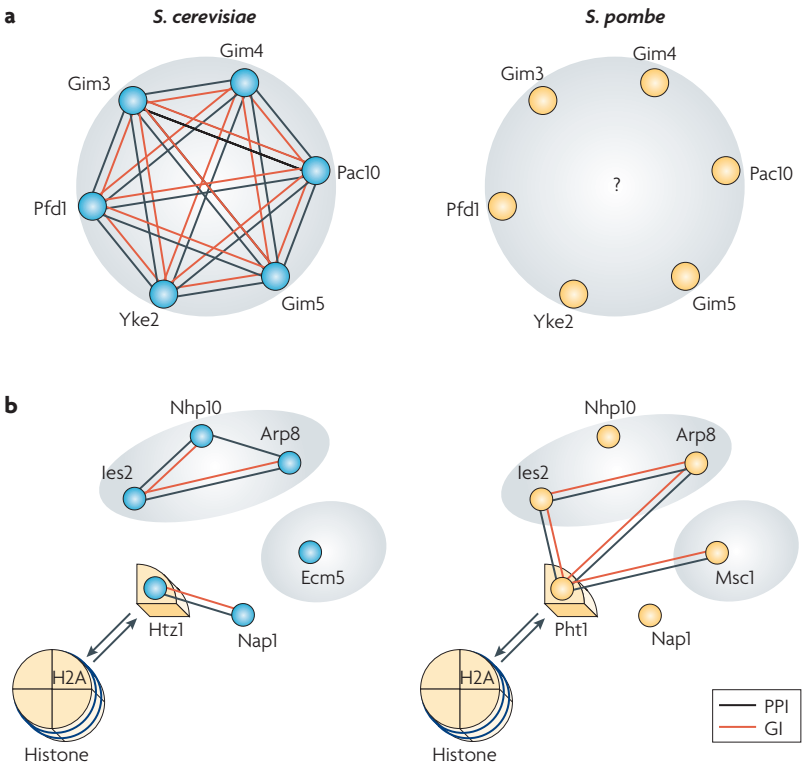


Figure 2 | **Network rewiring in *S. pombe* and *S. cerevisiae*.** A set of 234 single-copy orthologues[70,109], all of the links between which (from both genetic interactions (GIs) and protein–protein interactions (PPIs)) have been reported in the literature, was selected and changes in their interactions were considered as rewiring. Here, two examples are extracted from this network. The requirement for both PPIs and GIs was necessary to avoid misinterpretation due to limited coverage and accuracy of either network alone (for accuracy estimates see Supplementary information S2 (box)). **a** | Possible link losses within a network module in *Saccharomyces cerevisiae* (in which blue nodes represent proteins) and in *Schizosaccharomyces pombe* (in which yellow nodes represent proteins). The chaperone complex prefoldin is composed of the subunits genes involved in microtubule biogenesis 3 (Gim3), Gim4, Gim5, protein required in the absence of chromosome instability 10 (Pac10), prefoldin 1 (Pfd1) and yeast orthologue of mouse K-region expressed gene 2 (Yke2)[110]. This complex is conserved in eukaryotes and archaea, but none of the interactions between the subunits could be retrieved in the large-scale screen[70], nor is there other experimental support for interactions among the subunits in *S. pombe*. **b** | A rewiring example around the *S. cerevisiae* histone H2A variant Htz1 (the orthologue of *S. pombe* Pht1). Htz1 (or Pht1) can be exchanged with canonical histones (depicted as quarters of a circle) by the chromatin remodelling complexes inositol requiring protein 80 (INO80, which includes non-histone protein 10 (Nhp10), actin-related protein 8 (Arp8) and INO80 subunit 1 (Ies1)) and SWR1 (which includes the subunit extracellular matrix protein 5 (Ecm5) in *S. cerevisiae* and its orthologue, multicopy suppressor of Chk1 protein 1 (Msc1), in *S. pombe*). Htz1 also interacts with the chaperone nucleosome assembly protein 1 (Nap1) exclusively in *S. cerevisiae*. Nap1 also enables the exchange of Htz1 with canonical histone variants.

Evolution of links (topology)

The evolution of network links is coupled to that of nodes but is much more fine-tuned as links can change over time even if nodes are unaffected. The nature of the links determines the topology of a network and in this section we discuss the evolutionary dynamics of the links, highlight some constraints on link evolution and discuss some fundamental differences between PPI and metabolic networks in terms of topology and link evolution.

Evolutionary dynamics of links. Whereas network nodes (proteins) can be studied just by genome comparisons, the change of links and general topology requires sufficient network data in several species — data which are only slowly emerging. Based on incomplete yeast two-hybrid data it has been suggested that the rewiring of links can occur without gene duplication and that link changes might occur more frequently than node changes[38,52], although the limited accuracy and coverage of the networks hampers proper analysis (see Supplementary information S2 (box)).

Athough quantification of link changes remains difficult, there are plenty of genetic mechanisms that can easily lead to a link addition or deletion. Apart from point mutations, alternative splicing and domain accretion, inversion, shuffling and duplication are other means for the fast acquisition or loss of links. Combinations of these can even occur, whereby binding sites (forming the links) are under positive selection[68]. In fact, positive selection on proteins that are located peripherally in the interaction network has been reported[69], which is a more global indication of changing links.

We illustrate here the rewiring of links without node changes using large-scale data (FIG. 2). Unfortunately, to date there is limited data of sufficient quality to quantify these changes, but based on genetic interaction (GI) networks in *S. cerevisiae* and *Schizosaccharomyces pombe*, generated by double mutation, examples of rewiring of conserved functional modules have been described[70]. The respective data sets can also be used to project the rewiring to the node and link level, considering only those links that are confirmed by both GIs and PPIs and only those nodes that have a clear 1:1 orthologous relationship between the two species. These restrictions are needed to overcome sensitivity and selectivity issues in both networks, which could inflate the results. Although only a few in number, some rewiring events could be identified.

For example, the prefoldin complex is a highly connected protein cluster that is supported by GIs and PPIs in *S. cerevisiae*, whereas almost none of the respective interactions have been found between the orthologous proteins in *S. pombe* (FIG. 2a). The proteins shown in FIG. 2a form part of a multi-subunit chaperone complex that delivers unfolded proteins to a cytosolic chaperonin, which assists in their correct folding (see the *Saccharomyces* Genome Database). The functional characterization of the proteins constituting the prefoldin complex has been carried out in *S. cerevisiae*, and the functions of the orthologous proteins in *S. pombe* have been assumed based on these results. Given the lack of interactions in *S. pombe*, it is likely that some of the functionality of these proteins has changed after the divergence of the two yeasts.

A more complex example is the rewiring of the *S. cerevisiae* histone H2A variant Htz1 (the orthologue of *S. pombe* Pht1). Histone variants replace the canonical histone subunits and thereby change chromatin structure[71–73]. Histone replacement requires an interaction with chromatin remodelling complexes such as the inositol requiring protein 80 (INO80) complex, which facilitates the replacement of the canonical histone with the histone

variant Pht1 in *S. pombe* through INO80 subunit 2 (Ies2) and actin-related protein 8 (Arp8). In *S. cerevisiae*, the INO80 complex replaces the canonical histone with a different histone variant[71], therefore this part of the rewiring scenario in FIG. 2b has been known already. Furthermore, in both *S. pombe* and *S. cereviseae*, Htz1 also interacts with another chromatin remodelling complex, SWR1. Whereas in *S. pombe* this interaction seems to be mediated by multicopy suppressor of Chk1 protein 1 (Msc1), in *S. cerevisiae* this does not seem to be the case. Finally, another link addition or loss contributes to the rewiring observed in FIG. 2b: the interaction of Htz1 with the histone chaperone nucleosome assembly protein 1 (Nap1) in *S. cereviseae,* but not in *S. pombe*. Nap1 enables the replacement of Htz1 with canonical histone variants and vice versa by sliding the nucleosome[74–76]. Nap1 seems to have a different histone specificity in *S. pombe*[77]. This complex rewiring scenario is consistent with diverse remodelling and modification mechanisms for chromatin structures across different organisms[72].

The examples above illustrate that link rewiring can occur without node changes; the loss or gain of interactions (link loss or addition) might be a frequent event leading to the diversification of networks and can sometimes result in rewiring. A proper quantification of these events would require highly accurate and more complete networks. Even in the best studied PPI networks of yeast, there should be at least 1.5 times as many PPIs than are currently available[50]. Although network comparison only identifies rewiring events, it is likely that any link addition will be followed by a link loss, or vice versa, which means that rewiring is a result of two subsequent link-modifying steps.

Constraints on link addition and loss. As node addition and loss seem to have the largest impact on network topology, many constraints on links are coupled to constraints on the nodes. For example, in the yeast PPI network, central nodes (that is, nodes that are located in the central part of the network as defined by 'betweenness'; BOX 2) are encoded by essential genes[62]. This positive correlation between the essentiality and centrality of proteins is conserved in many species[78], suggesting that network features constrain the evolution of both proteins and their links. The node degree is also positively correlated with centrality, consistent with the classic proposal that more central proteins in PPI networks have more pleiotropic effects on cellular functions and, therefore, might be more constrained during evolution[79]. These constraints equally apply to links between central nodes, which tend to be essential. Constraints on link evolution are considerably different between PPI and metabolic networks. To understand these, their topological differences have to be discussed.

Different topologies in PPI and metabolic networks. Although PPI and metabolic networks overlap, a direct comparison is not simple and only a few studies have looked for topological differences between them (as quantifiable by the measures in BOX 1).

In the yeast PPI network, links between highly connected proteins are systematically suppressed, whereas those between highly connected and loosely connected pairs of proteins are favoured[80]. In metabolic networks the opposite has been observed[81]: interactors do have a similar degree (k) — the affected nodes have a similar number of links (FIG. 3). A recent report also shows that hub-like nodes tend to link low-degree nodes in PPI networks, whereas in metabolic networks interacting nodes have similar degrees[18] (FIG. 3). Together these data imply that PPI and metabolic networks have a fundamentally different network topology.

In addition, the assortativity of nodes, which is defined as the average connectivity of the nearest neighbours of a node (BOX 2) and which describes the density of the network module around the node, shows different trends in PPI and metabolic networks[18,80,81]. Connectivity and assortativity have a negative correlation in PPI networks but not in metabolic networks (FIG. 3). These observations reflect the nature of the networks. In PPI networks there are a few so-called super-hubs (for example, proteins such as chaperones that bind almost all other proteins) but the majority of proteins in PPI networks have only a few interaction partners, which means that they have a low degree. In metabolic networks most of the interacting proteins have similar degrees (FIG. 3). As a result, although there are also hubs in metabolic networks, these are not indiscriminate (they are not linked to nodes with low degrees). Instead, the interactors of hubs tend to also interact with each other. These tendencies in metabolic networks were observed in enzyme-centric networks (in which nodes are the proteins) and metabolite-centric networks (in which nodes are small molecules)[18].

PPI networks are less modular than metabolic networks. In general, a module of a biological network consists of a set of nodes that form a highly connected, coherent structural subsystem (for example, a protein complex) with a distinct function[82,83]. There are many of these modules in biological networks[61,84–87].

Several groups studied the evolution of these tightly connected modules. For example, using the conservation in genome organization (such as conserved gene neighbourhood) across many bacteria, numerous functional modules in *E. coli* were identified[88]. Similarly, phylogenetic profile similarity revealed that metabolic modules were conserved in many species[89]. However, only approximately half of the modules seem to be cohesive (conserved in evolution[86]) and even those can be subject to change[90]. Early on in network analysis, the modularity of metabolic networks was quantified. For example, when the average of cluster coefficients of a network ($C(n)$) was regarded as a modularity measure, all the studied metabolic networks of 43 distinct organisms were organized into highly connected topological modules[35]. Analysis of the more coherent modules revealed that ancient modules have been in essential processes such as translation and that there were also recently evolved modules that interact with the environment and are often horizontally transferred between species[90], implying a considerable impact on link evolution.

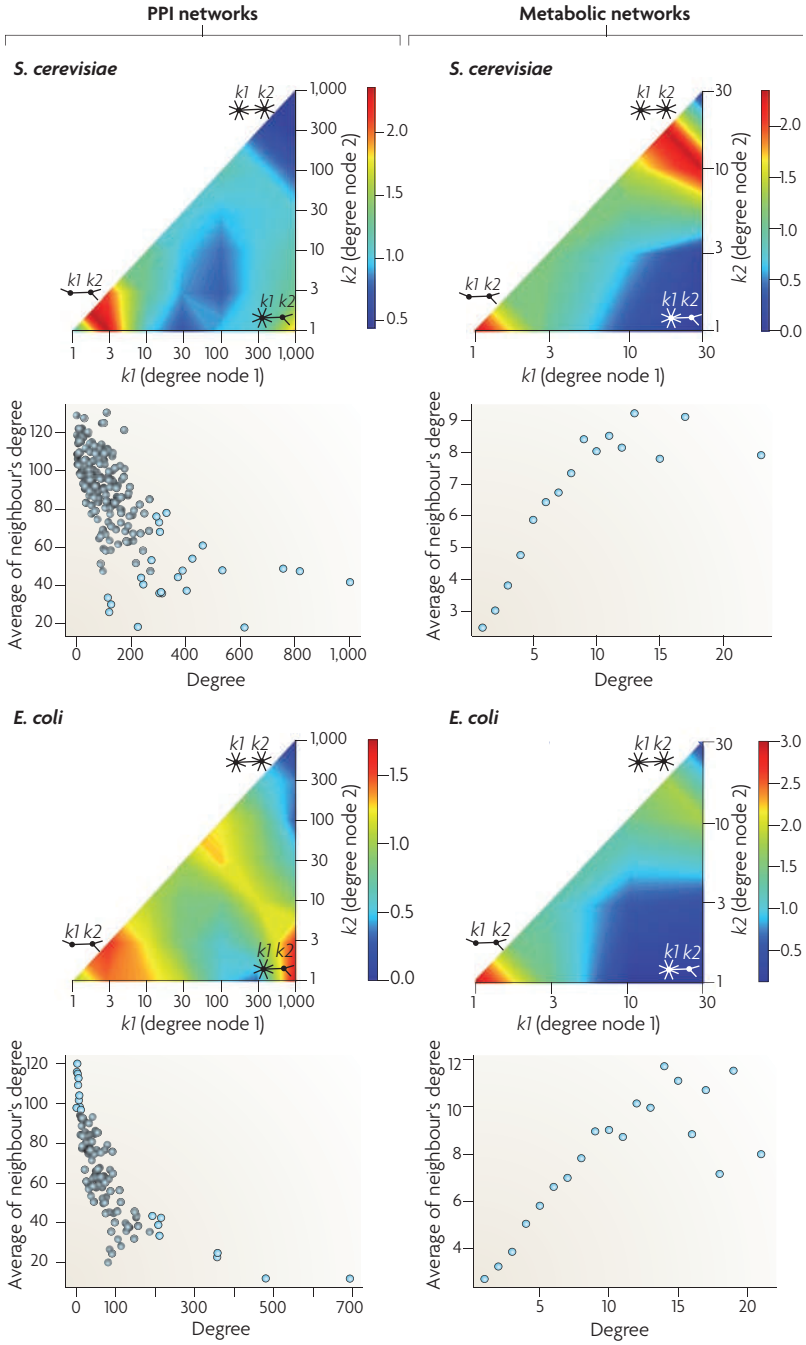


Figure 3 | **Topologies of PPI and metabolic networks in yeast and *E. coli*.** Both axes of the coloured plots (*k1* and *k2*) represent the degrees (or connectivity) of interacting proteins (nodes). A lower left position in the plot corresponds to interactions between two non-highly connected proteins, whereas upper right positions indicate hub–hub interactions (illustrated by schematic cartoons of connectivity). The colour gradient corresponds to the ratio *p*(*k1*,*k2*):*pr*(*k1*,*k2*), in which *p*(*k1*,*k2*) is the probability of an interacting pair with the respective node degrees *k1* and *k2*, and *pr*(*k1*,*k2*) is the same probability in a randomized network[80]. Thus, red indicates that the network has many interactions between non-hubs, whereas blue means that these types of interaction are avoided. The diagrams below the plots illustrate the assortativity (see also BOX 1) of networks. The horizontal axis is the connectivity of a node (degree), and the vertical axis is the average of the neighbour's connectivity. In cases where there is a negative correlation between these (as seen in protein–protein interaction (PPI) networks), nodes with high connectivity (hubs) tend to interact with nodes with relatively low connectivity. By contrast, a positive correlation (as seen in metabolic networks) suggests that hubs are usually located in highly connected topological modules.

Although most of these observations were inferred from an accumulation of interaction data and comparative genome analysis, the alignment of networks derived from *S. cerevisiae* and *H. pylori* directly revealed conserved functional modules (sub-networks) involving RNA polymerase and kinase signalling cascades and also many unexpected links between processes[91]. When applied to multiple species, more novel interactions and modules could be detected with statistical significance[24].

As metabolic modules are more cohesive than PPI modules[92], and as cohesive modules have a higher interconnectivity than non-cohesive ones[90], metabolic networks have a more modular structure than PPI networks, consistent with the differences in network structure described above. Modularity is also the basis for the hierarchy that is observed in biological networks. This has implications for the evolution of links, because links between modules are considerably less conserved and selection against link changes in modules should be particularly prominent in metabolic networks.

Genome size might have an impact on link evolution. A large-scale characterization of modularity in metabolic networks of more than 300 bacterial species revealed a positive correlation between modularity and genome size[83], suggesting that larger genomes have more modular metabolic networks and that network size is a strong determinant of metabolic modularity. Accordingly, small endosymbiotic organisms that tend to have smaller networks are less modular than non-symbiotic organisms. The modularity values of pathogens and commensals are generally even lower than those of endosymbionts. However, obligate mammalian pathogens that are transmitted by parasitic insect vectors are an exception as they have small networks that are highly modular[83]. This is supported by other reports which claim that the adaptation to different niches markedly enhances the evolution of modularity[93,94]. Although the relationship between modularity and genome size still has to be shown for PPI networks, it is becoming clear that there are various constraints on the evolution of links, mostly imposed by the modular structure of networks and the way environmental conditions affect this topology.

Towards 3D and 4D networks

So far, most of the research on networks has been devoted to *in vitro* and static networks, and these are usually considered in two dimensions (2D networks) without spatial (3D) or temporal (4D) resolution. In this section we discuss the 3D and 4D aspects of network evolution. Many network features and their evolution can be understood only when taking spatiotemporal resolution into account.

Spatiotemporal PPI networks. PPIs will never all happen at the same time and in the same place owing to the spatial or temporal separation of participating proteins[29,95]. In fact, PPI networks change considerably during the cell cycle or other dynamic processes[96], often within minutes. Even though the gene expression

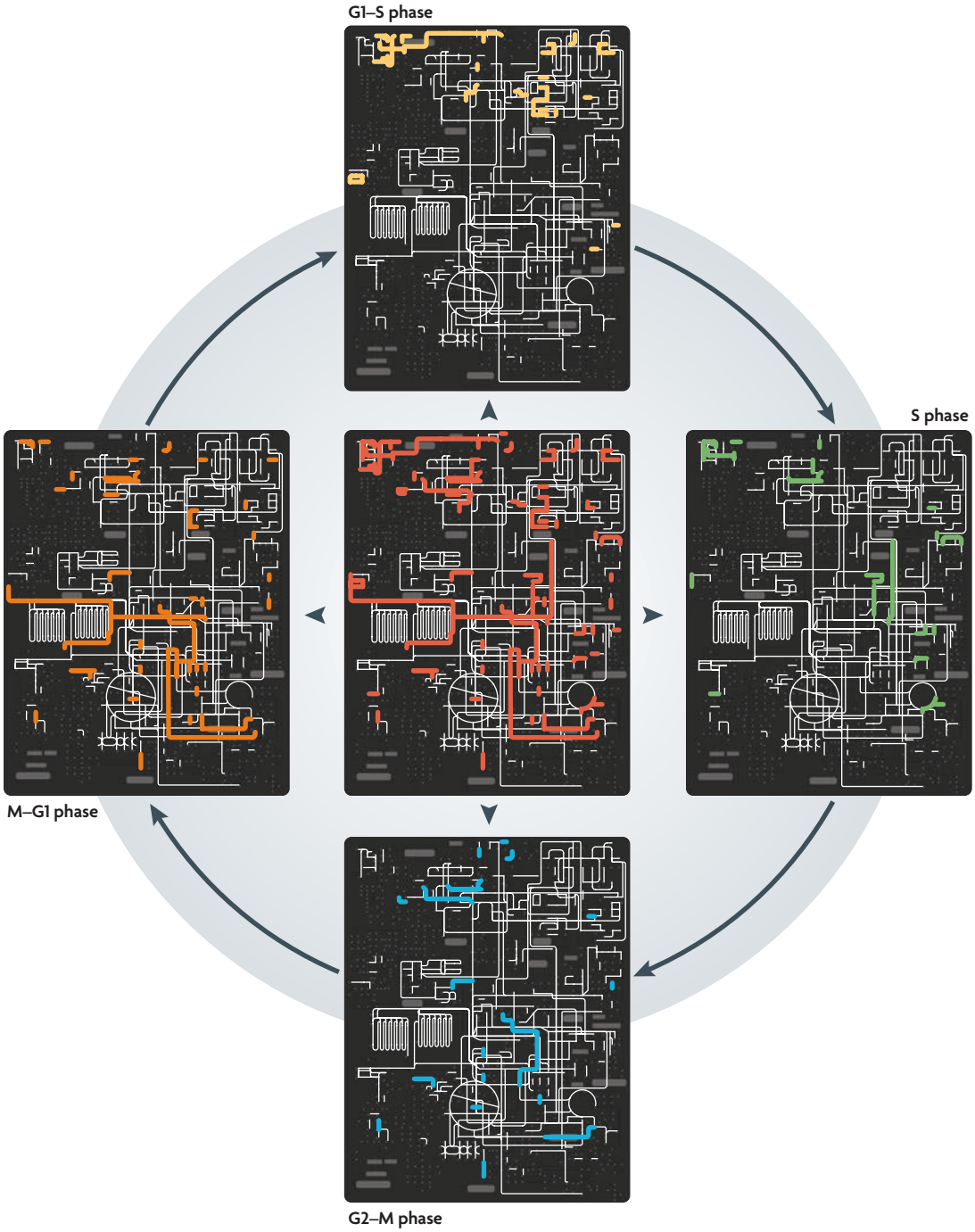


Figure 4 | **Metabolic dynamics during the yeast cell cycle.** In the central map, the currently known metabolic network of *Saccharomyces cerevisiae* is highlighted by white and red lines (corresponding to constitutive enzymatic reactions and periodical enzymatic reactions, respectively) above a black background of all of the metabolic reactions in the KEGG database[102]. The subset of metabolic enzymes that are periodically expressed during the cell cycle (160 enzymes out of a total of 600 periodic yeast genes[97]) is time-resolved in four distinct cell cycle phases (M–G1, orange lines; G1–S, yellow lines; S, green lines and G2–M, blue lines). In each phase, considerably different network structures can be observed despite a large fraction of constitutively expressed enzymes.

profile at a given time point in the cell cycle gives limited information about the status of the respective functional modules[97], PPI networks are highly dynamic. Time-dependent transcriptional regulation of individual genes influences the activation of complexes and hence their interaction behaviour[97]. The considerable gene expression differences of orthologues during the cell cycle of several eukaryotes coincide with differences in post-translational modification — transcriptional and post-translational regulation co-evolve[98].

The changes in post-translational modifications (such as phosphorylation sites), even in vertebrates, imply that novel interactions (through the kinase or phosphatase) can be introduced over short evolutionary timescales, independently of nodes.

Similarly, large temporal constraints on networks (and their evolution) are also evident from the coordinated gene expression in fly embryogenesis[99]. Here, in addition to individual cellular events, there are higher-order temporal processes that impose selection on network topology. Embryogenesis also implies specialization of cells and different spatial orientations. The complex gene expression patterns can be visualized by *in situ* hybridization of whole-mount embryos[100] or by monitoring whole organs, as has been carried out in *Aarabidopsis thaliana*[101]. Although these data have not exhaustively projected to biomolecular networks, resources on comparative 4D expression data, for example[102], indicate that PPI networks in each cell type are different and that there is a lot of variation between organisms.

Taken together, these studies show that there will be many constraints on the evolution of networks, indeed more than were previously expected, such as the cell type-specific gene expression patterns during an organism's lifetime. Existing groups of genes are switched off and on under various conditions and the network has to be robust enough to cope with these time- and cell type-specific modifications, such that not only the presence and absence of nodes in a genome but also the regulation of the respective genes have an impact on network evolution.

In addition, spatial aspects of a cell can affect network evolution. For example, different cell size or compartmental separation might lead to different fluxes and concentrations of metabolites, such that orthologues from two species might have to operate in a different network context.

Spatiotemporal metabolic networks. Although metabolic networks have been much less studied than PPI networks in terms of temporal and spatial aspects, it is clear that only sub-populations of metabolic proteins work together in particular tissues or cell types in multicellular organisms. For example, the human liver has the unique capacity to degrade ethanol and lipids and to detoxify various other compounds[103]. In fact, almost every cell type has a unique metabolic profile and, indeed, genome-wide data provide global support for tissue-specific metabolism in humans, for example[104,105]. Even within a cell, metabolic reactions are specialized in compartments, even more so in organelles such as mitochondria and peroxisomes, which have their own and sometimes independent metabolic systems. The localization of enzymes in compartments can vary between species, therefore the same (orthologous) enzymes will operate in different metabolic networks.

Network data with such spatial and temporal resolution will be needed to decipher where and when an interaction takes place[29]. To illustrate the temporal differences of metabolic networks, we re-analysed periodically expressed proteins during the yeast cell cycle (expression values taken from REF. 98) by focusing on metabolic proteins (FIG. 4). Each cell cycle stage has its own metabolic network and it is expected that different organisms will vary a lot in their expression profiles during the cell cycle[98]. Thus, a combined yeast metabolic network in 2D, as we know it from textbooks, captures only some of the network features. Network parameters might vary at different temporal conditions and selection might act on more subtle features such as the robustness towards concentration changes of substrates or regulatory aspects.

Perspectives

Networks and their evolution have been historically studied in different subfields, and distinctions between PPI, metabolic and transcriptional networks have been made. Although some of their properties differ (see above), a complete picture of the evolutionary constraints emerges only when they are viewed together and in conjunction with the interaction data on small molecules, lipids and carbohydrates that soon will also be available in large quantities. Nevertheless, it is clear already that PPI and metabolic networks are highly dynamic both during temporal processes such as the cell cycle and on evolutionary timescales. The nodes (proteins or enzymes) as well as the links (reactions or interactions) are continuously changing. Currently, high-quality interaction networks of reasonable coverage are available for only a small number of interaction types and model organisms, and even these might not have sufficient resolution to be able to reveal all the evolutionary trajectories. Instead of considering only general interactions, directionality, specifying actions such as phosphorylation events might need to be included. In the future, the main conceptual and methodological challenges will relate to the shift from static 2D to comparative spatiotemporal network analysis to reveal hitherto hidden constraints on the evolution of networks.

1. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
2. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
3. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574 (2001).
4. Rain, J. C. *et al.* The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).
5. Reboul, J. *et al. C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genet.* **34**, 35–41 (2003).
6. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
7. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).
8. Stelzl, U. *et al.* A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
9. Butland, G. *et al.* Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433**, 531–537 (2005).
10. Arifuzzaman, M. *et al.* Large-scale identification of protein–protein interaction of *Escherichia coli* K-12. *Genome Research* **16**, 686–691 (2006).
11. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
12. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
13. Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).

14. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
15. Tarassov, K. *et al.* An *in vivo* map of the yeast protein interactome. *Science* **320**, 1465–1470 (2008).
16. Fell, D. A. & Sauro, H. M. Metabolic control and its analysis. Additional relationships between elasticities and control coefficients. *Eur. J. Biochem.* **148**, 555–561 (1985).
17. Thomas, S. & Fell, D. A. A computer program for the algebraic determination of control coefficients in metabolic control analysis. *Biochem. J.* **292**, 351–360 (1993).
18. Durek, P. & Walther, D. The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles. *BMC systems biology* **2**, 100 (2008).
Provides the topological differences between PPI and metabolic networks.
19. Díaz-Mejía, J. J., Pérez-Rueda, E. & Segovia, L. A network perspective on the evolution of metabolism by gene duplication. *Genome Biology* **8**, R26 (2007).
20. Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol* **30**, 409–425 (1976).
21. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, 121 (2007).
22. Herrgård, M. J. *et al.* A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnol.* **26**, 1155–1160 (2008).
23. Duarte, N. *et al.* Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA* **104**, 1777–1782 (2007).
24. Sharan, R. *et al.* Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA* **102**, 1974–1979 (2005).
25. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).
26. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432 (2005).
27. Salwinski, L. *et al.* The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451 (2004).
28. Jensen, L. J. *et al.* STRING 8 — a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**, D412 (2009).
29. Jensen, L. J. & Bork, P. Biochemistry. Not comparable, but complementary. *Science* **322**, 56–57 (2008).
30. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
31. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
32. Bader, J. S., Chaudhuri, A., Rothberg, J. M. & Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnol.* **22**, 78–85 (2004).
33. Feist, A. M. & Palsson, B. Ø. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nature Biotechnol.* **26**, 659–667 (2008).
34. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
35. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
The first demonstration of a topological analysis for biomolecular networks, suggesting that the metabolic network is scale free.
36. Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).
37. van Noort, V., Snel, B. & Huynen, M. A. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* **5**, 280–284 (2004).
38. Wagner, A. How the global structure of protein interaction networks evolves. *Proc. Biol. Sci.* **270**, 457–466 (2003).
39. Rison, S. C. & Thornton, J. Pathway evolution, structurally speaking. *Current Opinion in Structural Biology* **12**, 374–382 (2002).
40. Janga, S. C. & Babu, M. M. Network-based approaches for linking metabolism with environment. *Genome Biology* **9**, 239 (2008).
41. Schmidt, S., Sunyaev, S., Bork, P. & Dandekar, T. Metabolites: a helping hand for pathway evolution? *Trends Biochem. Sci.* **28**, 336–341 (2003).
42. Horowitz, N. H. On the evolution of biochemical syntheses. *Proc. Natl Acad. Sci. USA* **31**, 153–157 (1945).
Together with reference 43, this paper provides the first evolutionary models of biochemical networks.
43. Ycas, M. On earlier states of the biochemical system. *J. Theor. Biol.* **44**, 145–160 (1974).
44. Lazcano, A. & Miller, S. L. On the origin of metabolic pathways. *J. Mol. Evol.* **49**, 424–431 (1999).
45. Copley, R. & Bork, P. Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627–641 (2000).
46. Teichmann, S. A. *et al.* The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J. Mol. Biol.* **311**, 693–708 (2001).
47. Alves, R., Chaleil, R. A. & Sternberg, M. J. Evolution of enzymes in metabolism: a network perspective. *J. Mol. Biol.* **320**, 751–770 (2002).
48. Raymond, J. & Segrè, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science* **311**, 1764–1767 (2006).
49. Borenstein, E., Kupiec, M., Feldman, M. W. & Ruppin, E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc. Natl Acad. Sci. USA* **105**, 14482–14487 (2008).
50. Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl Acad. Sci. USA* **106**, 1374–1379 (2009).
51. Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research* **12**, 17–25 (2002).
52. Berg, J., Lässig, M. & Wagner, A. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.* **4**, 51 (2004).
53. Campillos, M., Doerks, T., Shah, P. K. & Bork, P. Computational characterization of multiple Gag-like human proteins. *Trends Genet.* **22**, 585–589 (2006).
54. Liang, H. & Li, W. H. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* **23**, 375–378 (2007).
55. Rambaldi, D., Giorgi, F., Capuani, F., Ciliberto, A. & Ciccarelli, F. Low duplicability and network fragility of cancer genes. *Trends Genet.* **24**, 427–430 (2008).
56. Molina, N. & van Nimwegen, E. The evolution of domain-content in bacterial genomes. *Biology Direct* **3**, 51 (2008).
57. Maslov, S., Krishna, S., Pang, T. Y. & Sneppen, K. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proc. Natl Acad. Sci. USA* **106**, 9743–9748 (2009).
58. Raes, J., Korbel, J. O., Lercher, M. J., von Mering, C. & Bork, P. Prediction of effective genome size in metagenomic samples. *Genome Biology* **8**, R10 (2007).
59. Sorek, R. *et al.* Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* **318**, 1449 (2007).
60. Prachumwat, A. & Li, W. H. Protein function, connectivity, and duplicability in yeast. *Mol. Biol. Evo.* **23**, 30–39 (2006).
61. Han, J. D. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
62. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
First demonstration of a large-scale analysis of protein–protein physical interactions as a biomolecular network.
63. Wuchty, S. Evolution and topology in the yeast protein interaction network. *Genome Research* **14**, 1310–1314 (2004).
64. Fraser, H. B. Modularity and evolutionary constraint on proteins. *Nature Genet.* **37**, 351–352 (2005).
65. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evo.* **23**, 327–337 (2006).
66. Ekman, D. *et al.* What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? *Genome Biology* **7**, R45 (2006).
67. Lu, C. *et al.* Impacts of yeast metabolic network structure on enzyme evolution. *Genome Biology* **8**, 407 (2007).
68. Ciccarelli, F. *et al.* Complex genomic rearrangements lead to novel primate gene function. *Genome Research* **15**, 343–351 (2005).
69. Kim, P. M., Korbel, J. O. & Gerstein, M. B. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc. Natl Acad. Sci. USA* **104**, 20274–20279 (2007).
70. Roguev, A. *et al.* Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* **322**, 405–410 (2008).
Shows rewiring events on the genetic interaction network with large-scale experiments and analysis.
71. Conaway, R. & Conaway, J. The INO80 chromatin remodeling complex in transcription, replication and repair. *Trends Biochem. Sci.* **34**, 71–77 (2009).
72. Jin, J. *et al.* In and out: histone variant exchange in chromatin. *Trends Biochem. Sci.* **30**, 680–687 (2005).
73. Shevchenko, A. *et al.* Chromatin central: towards the comparative proteome by accurate mapping of the yeast proteomic environment. *Genome Biology* **9**, R167 (2008).
74. Lorch, Y., Zhang, M. & Kornberg, R. Histone octamer transfer by a chromatin-remodeling complex. *Cell* **96**, 389–392 (1999).
75. Park, Y., Chodaparambil, J. V., Bao, Y., McBryant, S. J. & Luger, K. Nucleosome assembly protein 1 exchanges histone H2A-H2B dimers and assists nucleosome sliding. *J. Biol. Chem.* **280**, 1817–1825 (2005).
76. Park, Y. J. & Luger, K. The structure of nucleosome assembly protein 1. *Proc. Natl Acad. Sci. USA* **103**, 1248–1253 (2006).
77. Walfridsson, J., Khorosjutina, O., Matikainen, P., Gustafsson, C. M. & Ekwall, K. A genome-wide role for CHD remodelling factors and Nap1 in nucleosome disassembly. *EMBO J.* **26**, 2868–2879 (2007).
78. Hahn, M. W. & Kern, A. D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evo.* **22**, 803–806 (2005).
79. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nature Rev. Genet.* **3**, 11–21 (2002).
80. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).
81. Zhu, D. & Qin, Z. S. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics* **6**, 8 (2005).
82. Wolf, D. M. & Arkin, A. P. Motifs, modules and games in bacteria. *Curr. Opin. Microbiol.* **6**, 125–134 (2003).
83. Kreimer, A., Borenstein, E., Gophna, U. & Ruppin, E. The evolution of modularity in bacterial metabolic networks. *Proc. Natl Acad. Sci. USA* **105**, 6976–6981 (2008).
84. Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA* **100**, 12123–12128 (2003).
85. Spirin, V., Gelfand, M. S., Mironov, A. A. & Mirny, L. A. A metabolic network in the evolutionary context: Multiscale structure and modularity. *Proc. Natl Acad. Sci. USA* **103**, 8774–8779 (2006).
86. Snel, B. & Huynen, M. A. Quantifying modularity in the evolution of biomolecular systems. *Genome Research* **14**, 391–397 (2004).
87. Ihmels, J., Levy, R. & Barkai, N. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nature Biotechnol.* **22**, 86–92 (2004).
88. von Mering, C. *et al.* Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA* **100**, 15428–15433 (2003).
89. Yamada, T., Kanehisa, M. & Goto, S. Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinformatics* **7**, 130 (2006).
90. Campillos, M., von Mering, C., Jensen, L. J. & Bork, P. Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Research* **16**, 374–382 (2006).
91. Kelley, B. P. *et al.* Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA* **100**, 11394–11399 (2003).
92. Fokkens, L. & Snel, B. Cohesive versus flexible evolution of functional modules in eukaryotes. *PLoS Comput. Biol.* **5**, e1000276 (2009).

93. Parter, M., Kashtan, N. & Alon, U. Environmental variability and modularity of bacterial metabolic networks. *BMC Evol. Biol.* **7**, 169 (2007).
94. Kashtan, N. & Alon, U. Spontaneous evolution of modularity and network motifs. *Proc. Natl Acad. Sci. USA* **102**, 13773–13778 (2005).
95. Bork, P. & Serrano, L. Towards Cellular Systems in 4D. *Cell* **121**, 507–509 (2005).
96. Laub, M. T., McAdams, H. H., Feldblyum, T., Fraser, C. M. & Shapiro, L. Global analysis of the genetic network controlling a bacterial cell cycle. *Science* **290**, 2144–2148 (2000).
97. de Lichtenberg, U., Jensen, L. J., Brunak, S. & Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724–727 (2005).
Provides a time-dependent protein interaction network by gene expression, leading to the study of protein complex dynamics.
98. Jensen, L. J., Jensen, T. S., de Lichtenberg, U., Brunak, S. & Bork, P. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **443**, 594–597 (2006).
99. Hooper, S. D. *et al.* Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis. *Mol. Syst. Biol.* **3**, 72 (2007).
100. Tomancak, P. *et al.* Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biology* **3**, 0088 (2002).
101. Schmid, M. *et al.* A gene expression map of *Arabidopsis thaliana* development. *Nature Genet.* **37**, 501–506 (2005).
102. Haudry, Y. *et al.* 4DXpress: a database for cross-species expression pattern comparisons. *Nucleic Acids Res.* **36**, D847–D853 (2008).
103. Berg, J., Tymoczko J., Stryer L. & Clarke N. *Biochemistry* (W. H. Freeman).
104. Shyamsundar, R. *et al.* A DNA microarray survey of gene expression in normal human tissues. *Genome Biology* **6**, R22 (2005).
105. Saito-Hisaminato, A. *et al.* Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. *DNA Res.* **9**, 35–45 (2002).
106. Erdős, P. & Renyi, A. On the strength of connectedness of a random graph. *Acta Math. Hung.* **12**, 261–267 (1961).
107. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
108. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089 (2005).
109. Roguev, A., Wiren, M., Weissman, J. S. & Krogan, N. J. High-throughput genetic interaction mapping in the fission yeast *Schizosaccharomyces pombe*. *Nature Methods* **4**, 861–866 (2007).
110. Siegert, R., Leroux, M. R., Scheufler, C., Hartl, F. U. & Moarefi, I. Structure of the molecular chaperone prefoldin: unique interaction of multiple coiled coil tentacles with unfolded proteins. *Cell* **103**, 621–632 (2000).

Acknowledgements
The authors thank V. van Noort for supporting statistical analysis of the network, J. Muller for supporting gene annotation procedures and all members of the Bork group for critical discussions. The work in the author's laboratory is supported by the BMBF grant Neuronet (17282) and the EU grant Metahit (17286).

DATABASES

UniProtKB: http://www.uniprot.org
Arp8 | Htz1 | Ies2 | Msc1 | Nap1 | Pht1
BioCyc: http://biocyc.org
DIP: http://dip.doe-mbi.ucla.edu/dip
KEGG: http://www.genome.jp/kegg
Reactome: http://www.reactome.org
SGD: http://www.yeastgenome.org
STRING: http://string.embl.de

FURTHER INFORMATION

Peer Bork's homepage: http://www.embl-heidelberg.de/~bork/
BioGRID: http://www.thebiogrid.org
eggNOG: http://eggnog.embl.de
iPath: http://pathways.embl.de

SUPPLEMENTARY INFORMATION

See online article: S1 (box) | S2 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF