

# NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION

*Albert-László Barabási\** & *Zoltán N. Oltvai<sup>‡</sup>*

A key aim of postgenomic biomedical research is to systematically catalogue all molecules and their interactions within a living cell. There is a clear need to understand how these molecules and the interactions between them determine the function of this enormously complex machinery, both in isolation and when surrounded by other cells. Rapid advances in network biology indicate that cellular networks are governed by universal laws and offer a new conceptual framework that could potentially revolutionize our view of biology and disease pathologies in the twenty-first century.

#### PROTEIN CHIPS

Similar to cDNA microarrays, this evolving technology involves arraying a genomic set of proteins on a solid surface without denaturing them. The proteins are arrayed at a high enough density for the detection of activity, binding to lipids and so on.

Reductionism, which has dominated biological research for over a century, has provided a wealth of knowledge about individual cellular components and their functions. Despite its enormous success, it is increasingly clear that a discrete biological function can only rarely be attributed to an individual molecule. Instead, most biological characteristics arise from complex interactions between the cell's numerous constituents, such as proteins, DNA, RNA and small molecules<sup>1–8</sup>. Therefore, a key challenge for biology in the twenty-first century is to understand the structure and the dynamics of the complex intercellular web of interactions that contribute to the structure and function of a living cell.

The development of high-throughput data-collection techniques, as epitomized by the widespread use of microarrays, allows for the simultaneous interrogation of the status of a cell's components at any given time. In turn, new technology platforms, such as PROTEIN CHIPS or semi-automated YEAST TWO-HYBRID SCREENS, help to determine how and when these molecules interact with each other. Various types of interaction webs, or networks, (including protein–protein interaction, metabolic, signalling and transcription-regulatory networks) emerge from the sum of these interactions. None of these networks are independent, instead they form a 'network of networks' that is responsible for the behaviour of the cell. A major challenge of contemporary biology is to embark on an integrated theoretical and experimental

programme to map out, understand and model in quantifiable terms the topological and dynamic properties of the various networks that control the behaviour of the cell.

Help along the way is provided by the rapidly developing theory of complex networks that, in the past few years, has made advances towards uncovering the organizing principles that govern the formation and evolution of various complex technological and social networks<sup>9–12</sup>. This research is already making an impact on cell biology. It has led to the realization that the architectural features of molecular interaction networks within a cell are shared to a large degree by other complex systems, such as the Internet, computer chips and society. This unexpected universality indicates that similar laws may govern most complex networks in nature, which allows the expertise from large and well-mapped non-biological systems to be used to characterize the intricate interwoven relationships that govern cellular functions.

In this review, we show that the quantifiable tools of network theory offer unforeseen possibilities to understand the cell's internal organization and evolution, fundamentally altering our view of cell biology. The emerging results are forcing the realization that, notwithstanding the importance of individual molecules, cellular function is a contextual attribute of strict and quantifiable patterns of interactions between the myriad of cellular constituents. Although uncovering the generic organizing principles of cellular networks

\**Department of Physics,  
University of Notre Dame,  
Notre Dame, Indiana 46556,  
USA.*

<sup>‡</sup>*Department of Pathology,  
Northwestern University,  
Chicago, Illinois 60611,  
USA.  
e-mails: alb@nd.edu;  
zno008@northwestern.edu  
doi:10.1038/nrg1272*

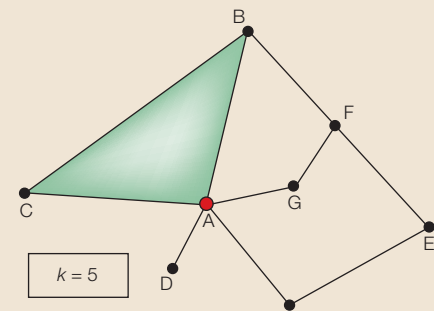
Box 1 | Network measures

Network biology offers a quantifiable description of the networks that characterize various biological systems. Here we define the most basic network measures that allow us to compare and characterize different complex networks.

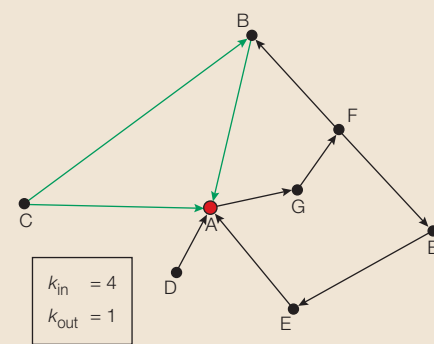
**Degree**

The most elementary characteristic of a node is its degree (or connectivity),  $k$ , which tells us how many links the node has to other nodes. For example, in the undirected network shown in part a of the figure, node A has degree  $k = 5$ . In networks in which each link has a selected direction (see figure, part b) there is an incoming degree,  $k_{in}$ , which denotes the number of links that point to a node, and an outgoing degree,  $k_{out}$ , which denotes the number of links that start from it. For example, node A in part b of the figure has  $k_{in} = 4$  and  $k_{out} = 1$ . An undirected network with  $N$  nodes and  $L$  links is characterized by an average degree  $\langle k \rangle = 2L/N$  (where  $\langle \rangle$  denotes the average).

**a Undirected network**



**b Directed network**



**Degree distribution**

The degree distribution,  $P(k)$ , gives the probability that a selected node has exactly  $k$  links.  $P(k)$  is obtained by counting the number of nodes  $N(k)$  with  $k = 1, 2 \dots$  links and dividing by the total number of nodes  $N$ . The degree distribution allows us to distinguish between different classes of networks. For example, a peaked degree distribution, as seen in a random network (BOX 2), indicates that the system has a characteristic degree and that there are no highly connected nodes (which are also known as hubs). By contrast, a power-law degree distribution indicates that a few hubs hold together numerous small nodes (BOX 2).

**Scale-free networks and the degree exponent**

Most biological networks are scale-free, which means that their degree distribution approximates a power law,  $P(k) \sim k^{-\gamma}$ , where  $\gamma$  is the degree exponent and  $\sim$  indicates 'proportional to'. The value of  $\gamma$  determines many properties of the system. The smaller the value of  $\gamma$ , the more important the role of the hubs is in the network. Whereas for  $\gamma > 3$  the hubs are not relevant, for  $2 > \gamma > 3$  there is a hierarchy of hubs, with the most connected hub being in contact with a small fraction of all nodes, and for  $\gamma = 2$  a hub-and-spoke network emerges, with the largest hub being in contact with a large fraction of all nodes. In general, the unusual properties of scale-free networks are valid only for  $\gamma < 3$ , when the dispersion of the  $P(k)$  distribution, which is defined as  $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$ , increases with the number of nodes (that is,  $\sigma$  diverges), resulting in a series of unexpected features, such as a high degree of robustness against accidental node failures<sup>71</sup>. For  $\gamma > 3$ , however, most unusual features are absent, and in many respects the scale-free network behaves like a random one.

**Shortest path and mean path length**

Distance in networks is measured with the path length, which tells us how many links we need to pass through to travel between two nodes. As there are many alternative paths between two nodes, the shortest path — the path with the smallest number of links between the selected nodes — has a special role. In directed networks, the distance  $\ell_{AB}$  from node A to node B is often different from the distance  $\ell_{BA}$  from B to A. For example, in part b of the figure,  $\ell_{AB} = 1$ , whereas  $\ell_{BA} = 3$ . Often there is no direct path between two nodes. As shown in part b of the figure, although there is a path from C to A, there is no path from A to C. The mean path length,  $\langle \ell \rangle$ , represents the average over the shortest paths between all pairs of nodes and offers a measure of a network's overall navigability.

**Clustering coefficient**

In many networks, if node A is connected to B, and B is connected to C, then it is highly probable that A also has a direct link to C. This phenomenon can be quantified using the clustering coefficient<sup>33</sup>  $C_i = 2n_i/k_i(k_i - 1)$ , where  $n_i$  is the number of links connecting the  $k_i$  neighbours of node  $i$  to each other. In other words,  $C_i$  gives the number of 'triangles' (see BOX 3) that go through node  $i$ , whereas  $k_i(k_i - 1)/2$  is the total number of triangles that could pass through node  $i$ , should all of node  $i$ 's neighbours be connected to each other. For example, only one pair of node A's five neighbours in part a of the figure are linked together (B and C), which gives  $n_A = 1$  and  $C_A = 2/20$ . By contrast, none of node F's neighbours link to each other, giving  $C_F = 0$ . The average clustering coefficient,  $\langle C \rangle$ , characterizes the overall tendency of nodes to form clusters or groups. An important measure of the network's structure is the function  $C(k)$ , which is defined as the average clustering coefficient of all nodes with  $k$  links. For many real networks  $C(k) \sim k^{-1}$ , which is an indication of a network's hierarchical character<sup>47,53</sup> (see BOX 2).

The average degree  $\langle k \rangle$ , average path length  $\langle \ell \rangle$  and average clustering coefficient  $\langle C \rangle$  depend on the number of nodes and links ( $N$  and  $L$ ) in the network. By contrast, the  $P(k)$  and  $C(k)$  functions are independent of the network's size and they therefore capture a network's generic features, which allows them to be used to classify various networks.

**YEAST TWO-HYBRID SCREEN**  
A genetic approach for the identification of potential protein–protein interactions. Protein X is fused to the site-specific DNA-binding domain of a transcription factor and protein Y to its transcriptional-activation domain — interaction between the proteins reconstitutes transcription-factor activity and leads to expression of reporter genes with recognition sites for the DNA-binding domain.

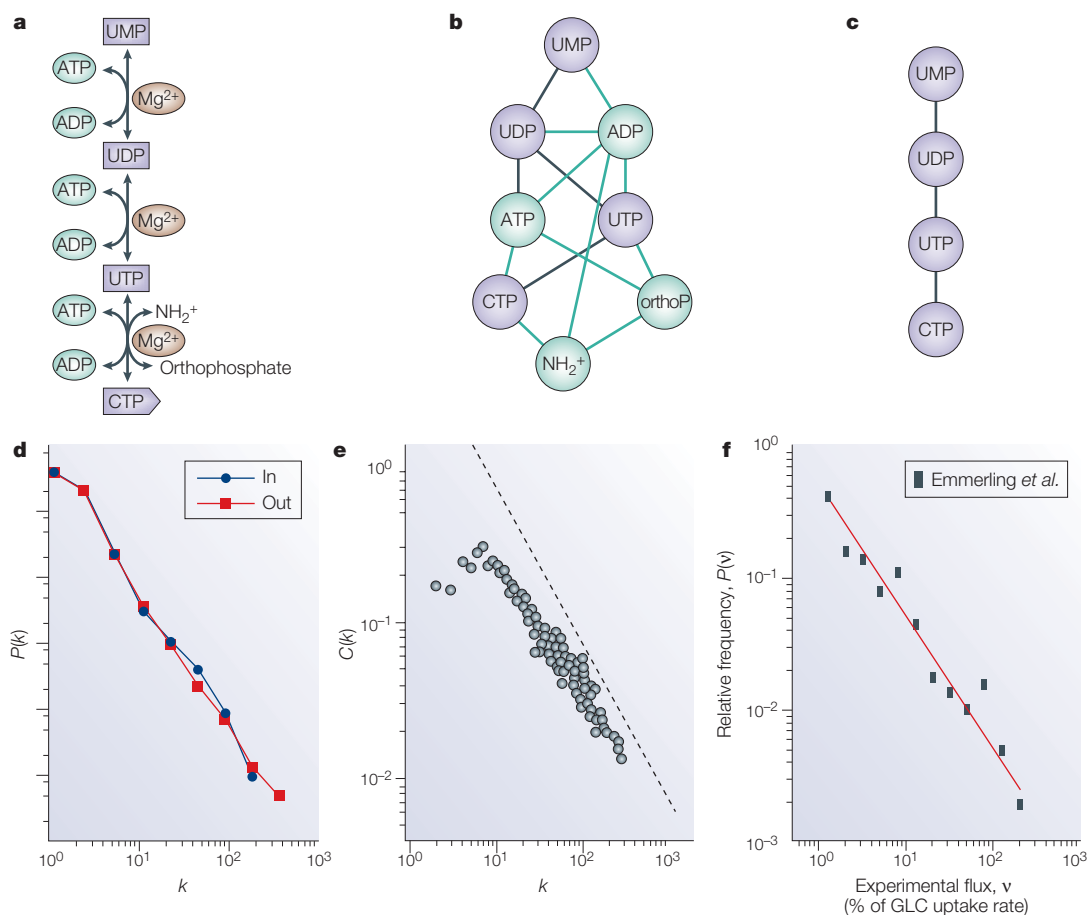
is fundamental to our understanding of the cell as a system, it also needs to develop relevance for the experimental biologist, helping to elucidate the role of individual molecules in various cellular processes. Therefore, we explore the specific biological details and the evolutionary origins that contribute to the formation of cellular networks, and the impact of the network structure on experimentally observable function and behavioural features. Our goal is to help understand the large-scale characteristics of cellular networks, complementing recent excellent reviews on the function of small genetic circuits (for example, see REFS 2,6). We also look to the future and the uncharted territories for which these approaches might bear further fruits.

**Basic network nomenclature**

The behaviour of most complex systems, from the cell to the Internet, emerges from the orchestrated activity of

many components that interact with each other through pairwise interactions. At a highly abstract level, the components can be reduced to a series of nodes that are connected to each other by links, with each link representing the interactions between two components. The nodes and links together form a network, or, in more formal mathematical language, a graph (BOX 1).

Establishing the identity of various cellular networks is not trivial. Physical interactions between molecules, such as protein–protein, protein–nucleic-acid and protein–metabolite interactions, can easily be conceptualized using the node-link nomenclature. Nevertheless, more complex functional interactions can also be considered within this representation. For example, small-molecule substrates can be envisioned as the nodes of a metabolic network and the links as the enzyme-catalysed reactions that transform one metabolite into another (FIG. 1a–c).



**Figure 1 | Characterizing metabolic networks.** To study the network characteristics of the metabolism a graph theoretic description needs to be established. Here, the graph theoretic description for a simple pathway (catalysed by  $Mg^{2+}$ -dependant enzymes) is illustrated (a). In the most abstract approach (b) all interacting metabolites are considered equally. The links between nodes represent reactions that interconvert one substrate into another. For many biological applications it is useful to ignore co-factors, such as the high-energy-phosphate donor ATP, which results in a second type of mapping (c) that connects only the main source metabolites to the main products. d | The degree distribution,  $P(k)$  of the metabolic network illustrates its scale-free topology<sup>16</sup>. e | The scaling of the clustering coefficient  $C(k)$  with the degree  $k$  illustrates the hierarchical architecture of metabolism<sup>53</sup> (The data shown in d and e represent an average over 43 organisms<sup>16,53</sup>). f | The flux distribution in the central metabolism of *Escherichia coli* follows a power law, which indicates that most reactions have small metabolic flux, whereas a few reactions, with high fluxes, carry most of the metabolic activity<sup>91</sup>. This plot is based on data that was collected by Emmerling *et al.*<sup>106</sup>. It should be noted that on all three plots the axis is logarithmic and a straight line on such log–log plots indicates a power-law scaling. CTP, cytidine triphosphate; GLC, aldo-hexose glucose; UDP, uridine diphosphate; UMP, uridine monophosphate; UTP, uridine triphosphate.

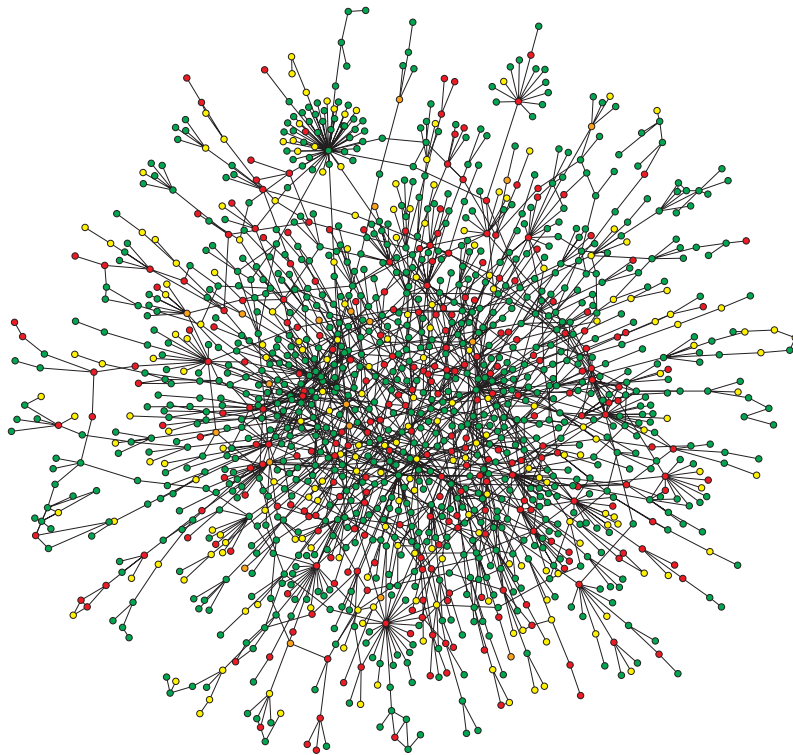


Figure 2 | **Yeast protein interaction network.** A map of protein–protein interactions<sup>18</sup> in *Saccharomyces cerevisiae*, which is based on early yeast two-hybrid measurements<sup>23</sup>, illustrates that a few highly connected nodes (which are also known as hubs) hold the network together. The largest cluster, which contains ~78% of all proteins, is shown. The colour of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown). Reproduced with permission from REF. 18 © Macmillan Magazines Ltd.

Depending on the nature of the interactions, networks can be directed or undirected. In directed networks, the interaction between any two nodes has a well-defined direction, which represents, for example, the direction of material flow from a substrate to a product in a metabolic reaction, or the direction of information flow from a transcription factor to the gene that it regulates. In undirected networks, the links do not have an assigned direction. For example, in protein interaction networks (FIG. 2) a link represents a mutual binding relationship: if protein A binds to protein B, then protein B also binds to protein A.

#### Architectural features of cellular networks

**From random to scale-free networks.** Probably the most important discovery of network theory was the realization that despite the remarkable diversity of networks in nature, their architecture is governed by a few simple principles that are common to most networks of major scientific and technological interest<sup>9,10</sup>. For decades graph theory — the field of mathematics that deals with the mathematical foundations of networks — modelled complex networks either as regular objects, such as a square or a diamond lattice, or as completely random network<sup>13</sup>. This approach was rooted in the influential work of two mathematicians, Paul Erdős, and Alfréd Rényi, who in 1960 initiated the study of the

mathematical properties of random networks<sup>14</sup>. Their much-investigated random network model assumes that a fixed number of nodes are connected randomly to each other (BOX 2). The most remarkable property of the model is its ‘democratic’ or uniform character, characterizing the degree, or connectivity ( $k$ ; BOX 1), of the individual nodes. Because, in the model, the links are placed randomly among the nodes, it is expected that some nodes collect only a few links whereas others collect many more. In a random network, the nodes degrees follow a Poisson distribution, which indicates that most nodes have roughly the same number of links, approximately equal to the network’s average degree,  $\langle k \rangle$  (where  $\langle \rangle$  denotes the average); nodes that have significantly more or less links than  $\langle k \rangle$  are absent or very rare (BOX 2).

Despite its elegance, a series of recent findings indicate that the random network model cannot explain the topological properties of real networks. The deviations from the random model have several key signatures, the most striking being the finding that, in contrast to the Poisson degree distribution, for many social and technological networks the number of nodes with a given degree follows a power law. That is, the probability that a chosen node has exactly  $k$  links follows  $P(k) \sim k^{-\gamma}$ , where  $\gamma$  is the degree exponent, with its value for most networks being between 2 and 3 (REF. 15). Networks that are characterized by a power-law degree distribution are highly non-uniform, most of the nodes have only a few links. A few nodes with a very large number of links, which are often called hubs, hold these nodes together. Networks with a power degree distribution are called scale-free<sup>15</sup>, a name that is rooted in statistical physics literature. It indicates the absence of a typical node in the network (one that could be used to characterize the rest of the nodes). This is in strong contrast to random networks, for which the degree of all nodes is in the vicinity of the average degree, which could be considered typical. However, scale-free networks could easily be called scale-rich as well, as their main feature is the coexistence of nodes of widely different degrees (scales), from nodes with one or two links to major hubs.

**Cellular networks are scale-free.** An important development in our understanding of the cellular network architecture was the finding that most networks within the cell approximate a scale-free topology. The first evidence came from the analysis of metabolism, in which the nodes are metabolites and the links represent enzyme-catalysed biochemical reactions (FIG. 1). As many of the reactions are irreversible, metabolic networks are directed. So, for each metabolite an ‘in’ and an ‘out’ degree (BOX 1) can be assigned that denotes the number of reactions that produce or consume it, respectively. The analysis of the metabolic networks of 43 different organisms from all three domains of life (eukaryotes, bacteria, and archaea) indicates that the cellular metabolism has a scale-free topology, in which most metabolic substrates participate in only one or two reactions, but a few, such as pyruvate or coenzyme A, participate in dozens and function as metabolic hubs<sup>16,17</sup>.

## Box 2 | Network models

Network models are crucial for shaping our understanding of complex networks and help to explain the origin of observed network characteristics. There are three models that had a direct impact on our understanding of biological networks.

**Random networks**

The Erdős–Rényi (ER) model of a random network<sup>14</sup> (see figure, part A) starts with  $N$  nodes and connects each pair of nodes with probability  $p$ , which creates a graph with approximately  $pN(N-1)/2$  randomly placed links (see figure, part Aa). The node degrees follow a Poisson distribution (see figure, part Ab), which indicates that most nodes have approximately the same number of links (close to the average degree  $\langle k \rangle$ ). The tail (high  $k$  region) of the degree distribution  $P(k)$  decreases exponentially, which indicates that nodes that significantly deviate from the average are extremely rare. The clustering coefficient is independent of a node's degree, so  $C(k)$  appears as a horizontal line if plotted as a function of  $k$  (see figure, part Ac). The mean path length is proportional to the logarithm of the network size,  $l \sim \log N$ , which indicates that it is characterized by the small-world property.

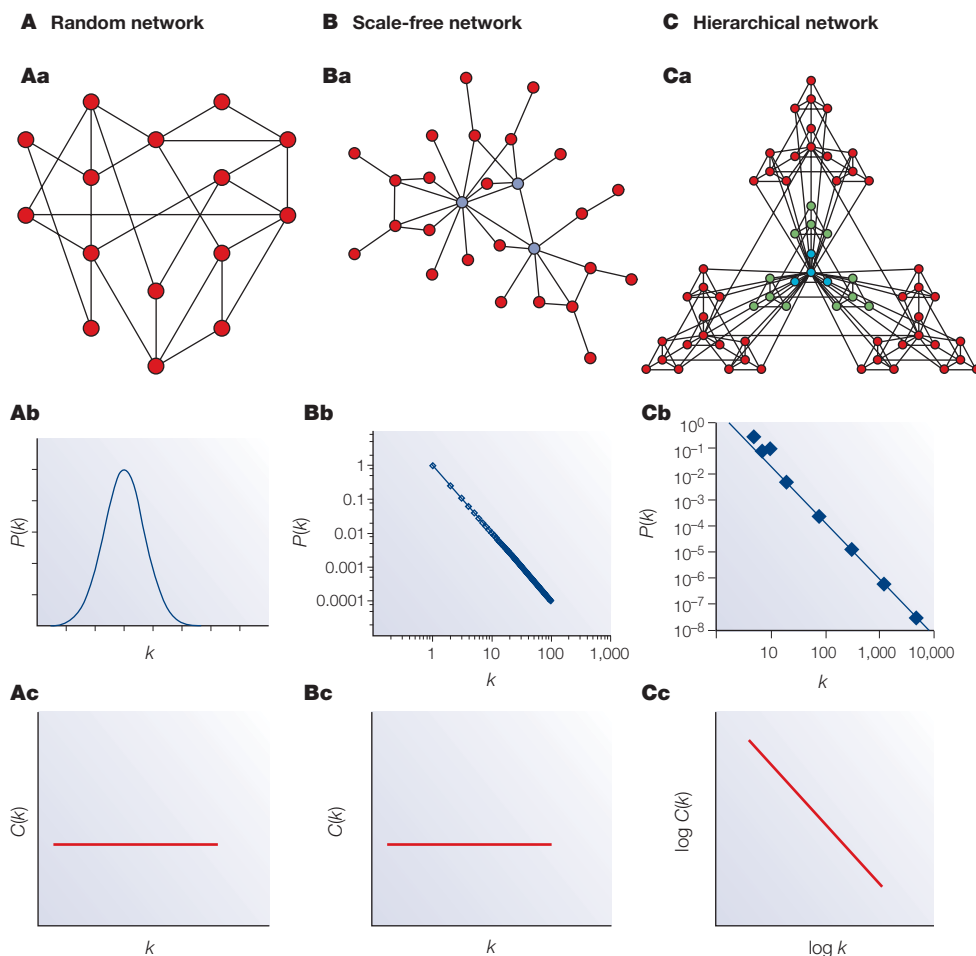
**Scale-free networks**

Scale-free networks (see figure, part B) are characterized by a power-law degree distribution; the probability that a node has  $k$  links follows  $P(k) \sim k^{-\gamma}$ , where  $\gamma$  is the degree exponent. The probability that a node is highly connected is statistically more significant than in a random graph, the network's properties often being determined by a relatively small number of highly connected nodes that are known as hubs (see figure, part Ba; blue nodes). In the Barabási–Albert model of a scale-free network<sup>13</sup>, at each time point a node with  $M$  links is added to the network, which connects to an already existing node  $I$  with probability  $\Pi_I = k_I/\sum k_j$ , where  $k_I$  is the degree of node  $I$  (FIG. 3) and  $J$  is the index denoting the sum over network nodes. The network that is generated by this growth process has a power-law degree distribution that is characterized by the degree exponent  $\gamma = 3$ . Such distributions are seen as a straight line on a log–log plot (see figure, part Bb). The network that is created by the Barabási–Albert model does not have an inherent modularity, so  $C(k)$  is independent of  $k$  (see figure, part Bc). Scale-free networks with degree exponents  $2 < \gamma < 3$ , a range that is observed in most biological and non-biological networks, are ultra-small<sup>34,35</sup>, with the average path length following  $l \sim \log \log N$ , which is significantly shorter than  $\log N$  that characterizes random small-world networks.

**Hierarchical networks**

To account for the coexistence of modularity, local clustering and scale-free topology in many real systems it has to be assumed that clusters combine in an iterative manner, generating a hierarchical network<sup>47,53</sup> (see figure, part C). The starting point of this construction is a small cluster of four densely linked nodes (see the four central nodes in figure, part Ca). Next, three replicas of this module are generated and the three external nodes of the replicated clusters

connected to the central node of the old cluster, which produces a large 16-node module. Three replicas of this 16-node module are then generated and the 16 peripheral nodes connected to the central node of the old module, which produces a new module of 64 nodes. The hierarchical network model seamlessly integrates a scale-free topology with an inherent modular structure by generating a network that has a power-law degree distribution with degree exponent  $\gamma = 1 + \ln 4 / \ln 3 = 2.26$  (see figure, part Cb) and a large, system-size independent average clustering coefficient  $\langle C \rangle \sim 0.6$ . The most important signature of hierarchical modularity is the scaling of the clustering coefficient, which follows  $C(k) \sim k^{-1}$  a straight line of slope  $-1$  on a log–log plot (see figure, part Cc). A hierarchical architecture implies that sparsely connected nodes are part of highly clustered areas, with communication between the different highly clustered neighbourhoods being maintained by a few hubs (see figure, part Ca).



As for direct physical interactions, several recent publications indicate that protein–protein interactions in diverse eukaryotic species also have the features of a scale-free network<sup>18–22</sup>. This is apparent in FIG. 2, which shows the protein interaction map of the yeast *Saccharomyces cerevisiae* as predicted by systematic two-hybrid screens<sup>23,24</sup>. Whereas most proteins participate in only a few interactions, a few participate in dozens — a typical feature of scale-free networks.

Further examples of scale-free organization include genetic regulatory networks, in which the nodes are individual genes and the links are derived from the expression correlations that are based on microarray data<sup>25,26</sup>, or protein domain networks that are constructed on the basis of protein domain interactions<sup>27,28</sup>. However, not all networks within the cell are scale-free. For example, the transcription regulatory networks of *S. cerevisiae* and *Escherichia coli* offer an interesting example of mixed scale-free and exponential characteristics. Indeed, the distribution that captures how many different genes a transcription factor interacts with follows a power law, which is a signature of a scale-free network. This indicates that most transcription factors regulate only a few genes, but a few general transcription factors interact with many genes. However, the incoming degree distribution, which tells us how many different transcription factors interact with a given gene, is best approximated by an exponential, which indicates that most genes are regulated by one to three transcription factors<sup>29,30</sup>. So, the key message is the recognition that cellular networks have a disproportionate number of highly connected nodes. Although the mathematical definition of a scale-free network requires us to establish that the degree distribution follows a power law, which is difficult in networks with too few nodes, the presence of hubs seems to be a general feature of all cellular networks, from regulatory webs to the p53 module<sup>31</sup>. These hubs fundamentally determine the network's behaviour (see below).

**Small-world effect and assortativity.** A common feature of all complex networks is that any two nodes can be connected with a path of a few links only. This 'small-world effect', which was originally observed in a social study<sup>32</sup>, has been subsequently shown in several systems, from neural networks<sup>33</sup> to the World Wide Web. Although the small-world effect is a property of random networks, scale-free networks are ultra small<sup>34,35</sup> — their path length is much shorter than predicted by the small-world effect (BOX 2). Within the cell, this ultra-small-world effect was first documented for metabolism, where paths of only three to four reactions can link most pairs of metabolites<sup>16,17</sup>. This short path length indicates that local perturbations in metabolite concentrations could reach the whole network very quickly. Interestingly, the evolutionarily reduced metabolic network of a parasitic bacterium has the same mean path length as the highly developed network of a large multicellular organism<sup>16</sup>, which indicates that there are evolutionary mechanisms that have maintained the average path length during evolution.

FIGURE 2 illustrates the disassortative nature of cellular networks. It indicates, for example, that, in protein interaction networks, highly connected nodes (hubs) avoid linking directly to each other and instead connect to proteins with only a few interactions<sup>36</sup>. In contrast to the assortative nature of social networks, in which well-connected people tend to know each other, disassortativity seems to be a property of all biological (metabolic, protein interaction) and technological (World Wide Web, Internet) networks<sup>36–38</sup>. Although the small- and ultra-small-world property of complex networks is mathematically well understood<sup>34,35</sup>, the origin of disassortativity in cellular networks remains unexplained.

**Evolutionary origin of scale-free networks.** The ubiquity of scale-free networks and hubs in technological, biological and social systems requires an explanation. It has emerged that two fundamental processes have a key role in the development of real networks<sup>15</sup>. First, most networks are the result of a growth process, during which new nodes join the system over an extended time period. This is the case for the World Wide Web, which has grown from 1 to more than 3-billion web pages over a 10-year period. Second, nodes prefer to connect to nodes that already have many links, a process that is known as preferential attachment. For example, on the World Wide Web we are more familiar with the highly connected web pages, and therefore are more likely to link to them. Growth and preferential attachment are jointly responsible for the emergence of the scale-free property in complex networks (FIG. 3a). Indeed, if a node has many links, new nodes will tend to connect to it with a higher probability. This node will therefore gain new links at a higher rate than its less connected peers and will turn into a hub<sup>15</sup>.

Growth and preferential attachment have a common origin in protein networks that is probably rooted in gene duplication<sup>39–44</sup>. Duplicated genes produce identical proteins that interact with the same protein partners (FIG. 3). Therefore, each protein that is in contact with a duplicated protein gains an extra link. Highly connected proteins have a natural advantage: it is not that they are more (or less) likely to be duplicated, but they are more likely to have a link to a duplicated protein than their weakly connected cousins, and therefore they are more likely to gain new links if a randomly selected protein is duplicated. This bias represents a subtle version of preferential attachment. The most important feature of this explanation is that it traces the origin of the scale-free topology back to a well-known biological mechanism — gene duplication. Although the role of gene duplication has been shown only for protein interaction networks, it probably explains, with appropriate adjustments, the emergence of the scale-free features in the regulatory and metabolic networks as well. It should be noted that, although the models show beyond doubt that gene duplication can lead to a scale-free topology, there is no direct proof that this mechanism is the only one, or the one that generates the observed power laws in cellular networks<sup>45</sup>. However, as gene duplication is a

major engineer of the genomic landscape, it is likely to be a key mechanism for generating the scale-free topology.

Two further results offer direct evidence that network growth is responsible for the observed topological features. The scale-free model (BOX 2) predicts that the nodes that appeared early in the history of the network are the most connected ones<sup>15</sup>. Indeed, an inspection of the metabolic hubs indicates that the remnants of the RNA world, such as coenzyme A, NAD and GTP, are among the most connected substrates of the metabolic network, as are elements of some of the most ancient metabolic pathways, such as glycolysis and the tricarboxylic acid cycle<sup>17</sup>. In the context of the protein interaction networks, cross-genome comparisons have found that, on average, the evolutionarily older proteins have more links to other proteins than their younger counterparts<sup>45,46</sup>. This offers direct empirical evidence for preferential attachment.

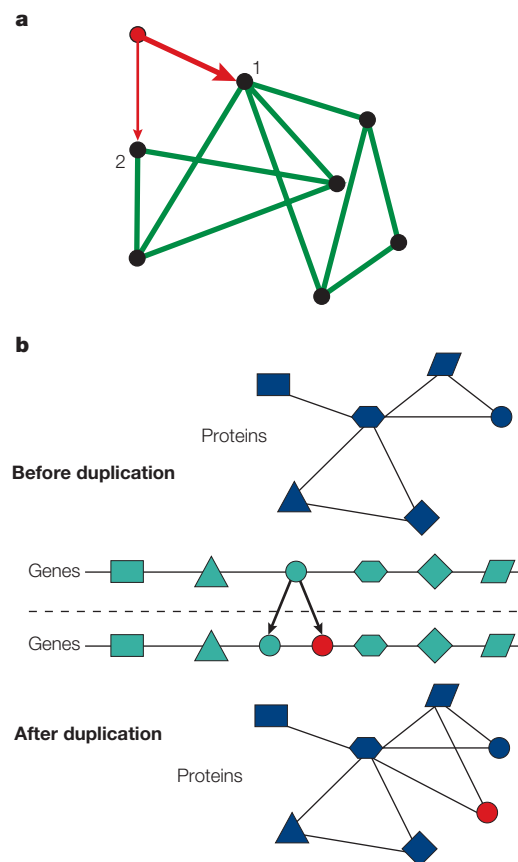
#### Motifs, modules and hierarchical networks

Cellular functions are likely to be carried out in a highly modular manner<sup>1</sup>. In general, modularity refers to a group of physically or functionally linked molecules (nodes) that work together to achieve a (relatively) distinct function<sup>1,6,8,47</sup>. Modules are seen in many systems, for example, circles of friends in social networks or websites that are devoted to similar topics on the World Wide Web. Similarly, in many complex engineered systems, from a modern aircraft to a computer chip, a highly modular structure is a fundamental design attribute.

Biology is full of examples of modularity. Relatively invariant protein–protein and protein–RNA complexes (physical modules) are at the core of many basic biological functions, from nucleic-acid synthesis to protein degradation<sup>48</sup>. Similarly, temporally coregulated groups of molecules are known to govern various stages of the cell cycle<sup>49–51</sup>, or to convey extracellular signals in bacterial chemotaxis or the yeast pheromone response pathway. In fact, most molecules in a cell are either part of an intracellular complex with modular activity, such as the ribosome, or they participate in an extended (functional) module as a temporally regulated element of a relatively distinct process (for example, signal amplification in a signalling pathway<sup>52</sup>).

To address the modularity of networks, tools and measures need to be developed that will allow us not only to establish if a network is modular, but also to explicitly identify the modules and their relationships in a given network.

**High clustering in cellular networks.** In a network representation, a module (or cluster) appears as a highly interconnected group of nodes. Each module can be reduced to a set of triangles (BOX 1); a high density of triangles is reflected by the clustering coefficient,  $C$  (REF. 33), the signature of a network's potential modularity (BOX 1). In the absence of modularity, the clustering coefficient of the real and the randomized network are comparable. The average clustering coefficient,  $\langle C \rangle$ , of



**Figure 3 | The origin of the scale-free topology and hubs in biological networks.** The origin of the scale-free topology in complex networks can be reduced to two basic mechanisms: growth and preferential attachment. Growth means that the network emerges through the subsequent addition of new nodes, such as the new red node that is added to the network that is shown in part **a**. Preferential attachment means that new nodes prefer to link to more connected nodes. For example, the probability that the red node will connect to node 1 is twice as large as connecting to node 2, as the degree of node 1 ( $k_1=4$ ) is twice the degree of node 2 ( $k_2=2$ ). Growth and preferential attachment generate hubs through a 'rich-gets-richer' mechanism: the more connected a node is, the more likely it is that new nodes will link to it, which allows the highly connected nodes to acquire new links faster than their less connected peers. In protein interaction networks, scale-free topology seems to have its origin in gene duplication. Part **b** shows a small protein interaction network (blue) and the genes that encode the proteins (green). When cells divide, occasionally one or several genes are copied twice into the offspring's genome (illustrated by the green and red circles). This induces growth in the protein interaction network because now we have an extra gene that encodes a new protein (red circle). The new protein has the same structure as the old one, so they both interact with the same proteins. Ultimately, the proteins that interacted with the original duplicated protein will each gain a new interaction to the new protein. Therefore proteins with a large number of interactions tend to gain links more often, as it is more likely that they interact with the protein that has been duplicated. This is a mechanism that generates preferential attachment in cellular networks. Indeed, in the example that is shown in part **b** it does not matter which gene is duplicated, the most connected central protein (hub) gains one interaction. In contrast, the square, which has only one link, gains a new link only if the hub is duplicated.

most real networks is significantly larger than that of a random network of equivalent size and degree distribution<sup>33</sup>. The metabolic network offers striking evidence for this:  $\langle C \rangle$  is independent of the network size, in contrast to a module-free scale-free network, for which  $\langle C \rangle$  decreases<sup>53</sup>. The cellular networks that have been studied so far, including protein interaction<sup>19</sup> and protein domain<sup>27</sup> networks, have a high  $\langle C \rangle$ , which indicates that high clustering is a generic feature of biological networks.

**Motifs are elementary units of cellular networks.** The high clustering indicates that networks are locally 'sprinkled' with various subgraphs of highly interlinked groups of nodes, which is a condition for the emergence of isolated functional modules. Subgraphs capture specific patterns of interconnections that characterize a given network at the local level (BOX 3). However, not all subgraphs are equally significant in real networks, as indicated by a series of recent observations<sup>29,30</sup>. To understand this, consider the highly regular square lattice: an inspection of its subgraphs would find very many squares and no triangles (BOX 3). It could (correctly) be concluded that the prevalence of squares and the absence of triangles tell us something fundamental about the architecture of a square lattice. In a complex network with an apparently random wiring diagram it is difficult to find such obvious signatures of order: all subgraphs, from triangles to squares or pentagons, are probably present. However, some subgraphs, which are known as motifs, are overrepresented when compared to a randomized version of the same network<sup>30,54</sup>. For example, triangle motifs, which are referred to as feed-forward loops (BOX 3) in directed networks, emerge in both transcription-regulatory and neural networks, whereas four-node feedback loops represent characteristic motifs in electric circuits but not in biological systems<sup>30</sup>. Each real network is characterized by its own set of distinct motifs, the identification of which provides information about the typical local interconnection patterns in the network<sup>30</sup>. The high degree of evolutionary conservation of motif constituents within the yeast protein interaction network<sup>35</sup> and the convergent evolution that is seen in the transcription-regulatory network of diverse species towards the same motif types<sup>56,57</sup> further indicate that motifs are indeed of direct biological relevance.

As the molecular components of a specific motif often interact with nodes that are outside the motif, how the different motifs interact with each other needs to be addressed. Empirical observations indicate that specific motif types aggregate to form large motif clusters. For example, in the *E. coli* transcription-regulatory network, most motifs overlap, generating distinct homologous motif clusters (BOX 3), in which the specific motifs are no longer clearly separable. As motifs are present in all of the real networks that have been examined so far<sup>30</sup>, it is likely that the aggregation of motifs into motif clusters is a general property of most real networks.

**Hierarchy organization of topological modules.** As the number of distinct subgraphs grows exponentially with the number of nodes that are in the subgraph, the study of larger motifs is combinatorially unfeasible. An alternative approach involves identifying groups of highly interconnected nodes, or modules, directly from the graph's topology and correlating these topological entities with their potential functional role. Module identification is complicated by the fact that at face value the scale-free property and modularity seem to be contradictory. Modules by definition imply that there are groups of nodes that are relatively isolated from the rest of the system. However, in a scale-free network hubs are in contact with a high fraction of nodes, which makes the existence of relatively isolated modules unlikely. Clustering and hubs naturally coexist, however, which indicates that topological modules are not independent, but combine to form a hierarchical network<sup>47,53</sup>.

An example of such a hierarchical network is shown in BOX 2; this network is simultaneously scale-free and has a high clustering coefficient that is independent of system size. The network is made of many small, highly integrated 4-node modules that are assembled into larger 16-node modules, each of which combines in a hierarchical fashion into even larger 64-node modules. The quantifiable signature of hierarchical modularity is the dependence of the clustering coefficient on the degree of a node, which follows  $C(k) \sim k^{-1}$  (REF 58). This indicates that nodes with only a few links have a high  $C$  and belong to highly interconnected small modules. By contrast, the highly connected hubs have a low  $C$ , with their role being to link different, and otherwise not communicating, modules. It should be noted that the random and scale-free models that are shown in BOX 2 do not have a hierarchical topology, because  $C(k)$  is independent of  $k$  in their case. This is not surprising, as their construction does not contain elements that would favour the emergence of modules.

**Identifying topological and functional modules.** Signatures of hierarchical modularity are present in all cellular networks that have been investigated so far, ranging from metabolic<sup>53</sup> to protein-protein interaction<sup>20,22</sup> and regulatory networks. But can the modules that are present in a cellular network be determined in an automated and objective fashion? This would require a unique breakdown of the cellular network into a set of biologically relevant functional modules. The good news is that if there are clearly separated modules in the system, most clustering methods can identify them. Indeed, several methods have recently been introduced to identify modules in various networks, using either the network's topological description<sup>53,59-64</sup> or combining the topology with integrated functional genomics data<sup>65-70</sup>. It must be kept in mind, however, that different methods predict different boundaries between modules that are not sharply separated. This ambiguity is not only a limitation of the present clustering methods, but it is a consequence of the network's hierarchical modularity.

The hierarchical modularity indicates that modules do not have a characteristic size: the network is as likely



to be partitioned into a set of clusters of 10–20 components (metabolites, genes) as into fewer, but larger modules. At present there are no objective mathematical criteria for deciding that one partition is better than another. Indeed, in most of the present clustering algorithms some internal parameter controls the typical size of the uncovered modules, and changing the parameter results in a different set of larger or smaller

modules. Does this mean that it is inherently impossible to identify the modules in a biological network? From a mathematical perspective it does indeed indicate that looking for a set of unique modules is an ill-defined problem. An easy solution, however, is to avoid seeking a breakdown into an absolute set of modules, but rather to visualize the hierarchical relationship between modules of different sizes<sup>53,61,62</sup>. The identification of the groups of

### Box 3 | Subgraphs, motifs and motif clusters

Whereas the scale-free and hierarchical features of complex networks emphasize the organizing principles that determine the network's large-scale structure, an alternative approach starts from the bottom and looks for highly representative patterns of interactions that characterize specific networks.

#### Subgraphs

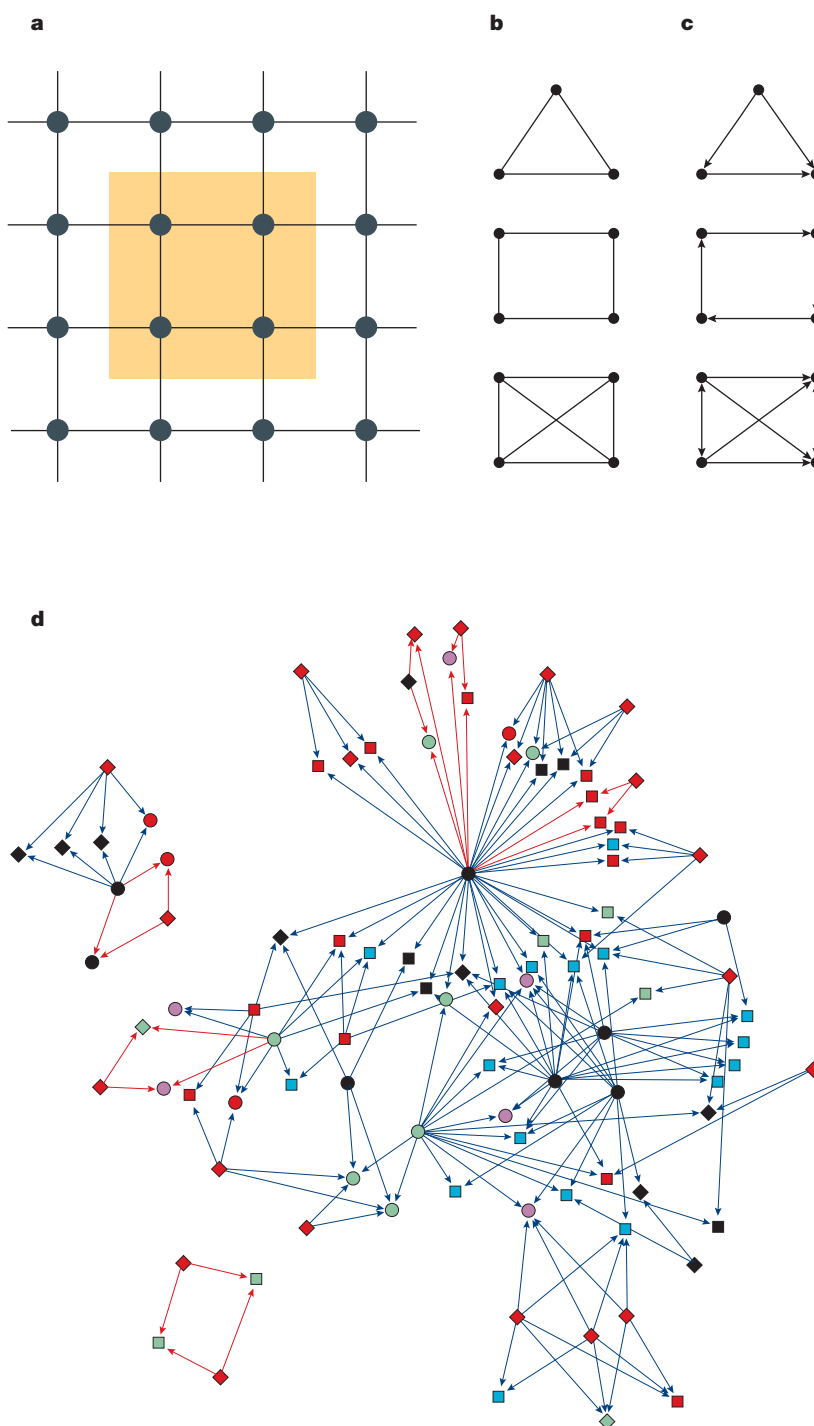
A connected subgraph represents a subset of nodes that are connected to each other in a specific wiring diagram. For example, in part a of the figure four nodes that form a little square (yellow) represent a subgraph of a square lattice. Networks with a more intricate wiring diagram can have various different subgraphs. For example, in part A of the figure in BOX 1, nodes A, B and C form a triangle subgraph, whereas A, B, F and G form a square subgraph. Examples of different potential subgraphs that are present in undirected networks are shown in part b of the figure (a directed network is shown in part c). It should be noted that the number of distinct subgraphs grows exponentially with an increasing number of nodes.

#### Motifs

Not all subgraphs occur with equal frequency. Indeed, the square lattice (see figure, part a) contains many squares, but no triangles. In a complex network with an apparently random wiring diagram all subgraphs — from triangles to squares and pentagons and so on — are present. However, some subgraphs, which are known as motifs, are over represented as compared to a randomized version of the same network<sup>30,54</sup>. For example, the directed triangle motif that is known as the feed-forward loop (see figure, top of part c) emerges in both transcription-regulatory and neural networks, whereas four-node feedback loops (see figure, middle of part c) represent characteristic motifs in electric circuits but not in biological systems<sup>30</sup>. To identify the motifs that characterize a given network, all subgraphs of  $n$  nodes in the network are determined. Next, the network is randomized while keeping the number of nodes, links and the degree distribution unchanged. Subgraphs that occur significantly more frequently in the real network, as compared to randomized one, are designated to be the motifs.

#### Motif clusters

The motifs and subgraphs that occur in a given network are not independent of each other. In part d of the figure, all of the 209 bi-fan motifs (a motif with 4 nodes) that are found in the *Escherichia coli* transcription-regulatory network<sup>29</sup> are shown simultaneously. As the figure shows, 208 of the 209 bi-fan motifs form two extended motif clusters (R. Dobrin *et al.*, manuscript in preparation) and only one motif remains isolated (bottom left corner). Such clustering of motifs into motif clusters seems to be a general property of all real networks. In part d of the figure the motifs that share links with other motifs are shown in blue; otherwise they are red. The different colours and shapes of the nodes illustrate their functional classification.



molecules of various sizes that together carry out a specific cellular function is a key issue in network biology, and one that is likely to witness much progress in the near future.

#### Network robustness

A key feature of many complex systems is their robustness, which refers to the system's ability to respond to changes in the external conditions or internal organization while maintaining relatively normal behaviour. To understand the cell's functional organization, insights into the interplay between the network structure and robustness, as well as their joint evolutionary origins, are needed.

**Topological robustness.** Intuition tells us that disabling a substantial number of nodes will result in an inevitable functional disintegration of a network. This is certainly true for a random network: if a critical fraction of nodes is removed, a phase transition is observed, breaking the network into tiny, non-communicating islands of nodes. Complex systems, from the cell to the Internet, can be amazingly resilient against component failure, withstanding even the incapacitation of many of their individual components and many changes in external conditions. We have recently learnt that topology has an important role in generating this topological robustness<sup>71</sup>. Scale-free networks do not have a critical threshold for disintegration — they are amazingly robust against accidental failures: even if 80% of randomly selected nodes fail, the remaining 20% still form a compact cluster with a path connecting any two nodes. This is because random failure affects mainly the numerous small degree nodes, the absence of which doesn't disrupt the network's integrity<sup>71</sup>. This reliance on hubs, on the other hand, induces a so-called attack vulnerability — the removal of a few key hubs splinters the system into small isolated node clusters<sup>71</sup>.

This double-edged feature of scale-free networks indicates that there is a strong relationship between the hub status of a molecule (for example, its number of links) and its role in maintaining the viability and/or growth of a cell. Deletion analyses indicate that in *S. cerevisiae* only ~10% of the proteins with less than 5 links are essential, but this fraction increases to over 60% for proteins with more than 15 interactions, which indicates that the protein's degree of connectedness has an important role in determining its deletion phenotype<sup>18</sup>. Furthermore, only ~18.7% of *S. cerevisiae* genes (~14.4% in *E. coli*) are lethal when deleted individually<sup>72–74</sup>, and the simultaneous deletion of many *E. coli* genes is without substantial phenotypic effect<sup>75,76</sup>. These results are in line with the expectation that many lightly connected nodes in a scale-free network do not have a major effect on the network's integrity. The importance of hubs is further corroborated by their evolutionary conservation: highly interacting *S. cerevisiae* proteins have a smaller evolutionary distance to their orthologues in *Caenorhabditis elegans*<sup>77</sup> and are more likely to have orthologues in higher organisms<sup>78</sup>.

**Functional and dynamical robustness.** A complete understanding of network robustness requires that the functional and dynamic changes that are caused by perturbations are explored. In a cellular network, each node has a slightly different biological function and therefore the effect of a perturbation cannot depend on the node's degree only. This is well illustrated by the finding that experimentally identified protein complexes tend to be composed of uniformly essential or non-essential molecules<sup>79</sup>. This indicates that the functional role (dispensability) of the whole complex determines the deletion phenotype of the individual proteins.

The functional and dynamical robustness of cellular networks is supported by recent results that indicate that several relatively well-delineated extended modules are robust to many varied perturbations. For example, the chemotaxis receptor module of *E. coli* maintains its normal function despite significant changes in a specified set of internal or external parameters, which leaves its tumbling frequency relatively unchanged even under orders-of-magnitude deviations in the rate constants or ligand concentrations<sup>80,81</sup>. The development of the correct segment polarity pattern in *Drosophila melanogaster* embryos is also robust to marked changes in the initial conditions, reaction parameters, or to the absence of certain gene products<sup>82,83</sup>. However, similar to topological robustness, dynamical and functional robustness are also selective: whereas some important parameters remain unchanged under perturbations, others vary widely. For example, the adaptation time or steady-state behaviour in chemotaxis show strong variations in response to changes in protein concentrations<sup>81</sup>.

Although our understanding of network robustness is far from complete, a few important themes have emerged. First, it is increasingly accepted that adaptation and robustness are inherent network properties, and not a result of the fine-tuning of a component's characteristics<sup>80,81</sup>. Second, robustness is inevitably accompanied by vulnerabilities: although many cellular networks are well adapted to compensate for the most common perturbations, they collapse when well-selected network components are disrupted. Third, the ability of a module to evolve also has a key role in developing or limiting robustness<sup>84</sup>. Indeed, evolutionarily 'frozen' modules that are responsible for key cellular functions, such as nucleic-acid synthesis, might be less able to withstand uncommon errors, such as the inactivation of two molecules within the same functional module. For example, orotate phosphoribosyltransferase (pyrE)-challenged *E. coli* cells cannot tolerate further gene inactivation in the evolutionarily highly conserved pyrimidine metabolic module, even in rich cultural media<sup>74</sup>. Finally, modularity and robustness are presumably considerably quite intertwined, with the weak communication between modules probably limiting the effects of local perturbations in cellular networks.

#### Beyond topology: characterizing the links

Despite their successes, purely topology-based approaches have important intrinsic limitations. For example, the activity of the various metabolic reactions or regulatory

interactions differs widely: some are highly active under most growth conditions, others switch on only under rare environmental circumstances. Therefore, an ultimate description of cellular networks requires that both the intensity (that is, strength) and the temporal aspects of the interactions are considered<sup>85,86</sup>. Although, so far, we know little about the temporal aspects of the various cellular interactions, recent results have shed light on how the strength of the interactions is organized in metabolic and genetic-regulatory networks.

In metabolic networks, the flux of a given metabolic reaction, which represents the amount of substrate that is being converted to a product within a unit of time, offers the best measure of interaction strength. Metabolic flux–balance approaches (FBA)<sup>87,88</sup>, which allow the flux for each reaction to be calculated, have recently significantly improved our ability to make quantifiable predictions on the relative importance of various reactions, giving rise to experimentally testable hypotheses<sup>89,90</sup>. A striking feature of the flux distribution of *E. coli* is its overall heterogeneity: reactions with flux that spans several orders of magnitude coexist under the same conditions. This is captured by the flux distribution for *E. coli*, which follows a power law. This indicates that most reactions have quite small fluxes, coexisting with a few reactions with extremely high flux values<sup>91</sup>.

A similar pattern is observed when the strength of the various genetic regulatory interactions that are provided by microarray datasets are investigated<sup>92</sup>. Capturing the degree to which each pair of genes is coexpressed (that is, assigning each pair a correlation coefficient) or examining the local similarities in perturbed transcriptome profiles of *S. cerevisiae* indicates that the functional organization of genetic regulatory networks might also be highly uneven<sup>93,94</sup>. That is, although most of them only have weak correlations, a few pairs show quite a significant correlation coefficient. These highly correlated pairs probably correspond to direct regulatory and protein interactions. This hypothesis is supported by the finding that the correlations are higher along the links of the protein interaction network or between proteins that occur in the same complex as compared to pairs of proteins that are not known to interact directly<sup>79,95–97</sup>.

Taken together, these results indicate that the biochemical activity in both the metabolic and genetic networks is dominated by several ‘hot links’ that represent high activity interactions that are embedded into a web of less active interactions. This attribute does not seem to be a unique feature of biological systems: there are hot links in many non-biological networks, their activity following a wide distribution<sup>98–100</sup>. The origin of this seemingly universal property of the links is probably rooted again in the network topology. Indeed, it seems that the metabolic fluxes and the weights of links in some non-biological systems are uniquely determined by the scale-free nature of the network topology<sup>91,98</sup>. At present, a more general principle that could explain the coexpression distribution data equally well is lacking.

### Future directions

Despite the significant advances in the past few years, (molecular) network biology is only in its infancy. Future progress is expected in many directions, ranging from the development of new theoretical methods to characterize the network topology to insights into the dynamics of motif clusters and biological function. Most importantly, to move significantly beyond our present level of knowledge, we need to enhance our data collection abilities. This will require the development of highly sensitive tools for identifying and quantifying the concentrations, fluxes and interactions of various types of molecules at high resolution both in space and time<sup>101</sup>. In the absence of such comprehensive data sets, whole arrays of functionally important cellular networks remain completely unexplored, ranging from signalling networks to the role of microRNAs in network topology and dynamics.

Similarly, most work at present focuses on the totality of interactions or snapshots of activity in a few selected environments and in an abstract space. However, a cell’s internal state or position in the cell cycle, for example, is a key determinant of actual interactions<sup>102</sup> that requires data collection in distinct functional and temporal states. Equally importantly, all these interactions take place in the context of the cell’s physical existence. So, its unique intracellular milieu, three-dimensional shape, anatomical architecture, compartmentalization and the state of its cytoskeleton are likely to further restrict the potential interactions in cellular networks. Finally, most studies have so far focused on different subsets of the complex cellular networks. Integrated studies that allow us to look at all (metabolic, regulatory, spatial and so on) interactions could offer further insights into how the network of networks contributes to the cell’s observable behaviour, as shown for the *S. cerevisiae* galactose utilization pathway<sup>103</sup>. Extending them to the whole cellular network of an organism is the ultimate aim of network and systems biology.

### Conclusions

It is impossible to ignore the apparent universality we have witnessed by delving into the totality of pairwise interactions among the various molecules of a cell. Instead of chance and randomness, we have found a high degree of internal order that governs the cell’s molecular organization. Along the way, a new language has been created, which allows the cell’s molecular makeup to be discussed as a network of interacting constituents, and to spot and quantify the interplay between behaviour, structure and function. The cell can be approached from the bottom up, moving from molecules to motifs and modules, or from the top to the bottom, starting from the network’s scale-free and hierarchical nature and moving to the organism-specific modules and molecules<sup>5</sup>. In either case, it must be acknowledged that structure, topology, network usage, robustness and function are deeply interlinked, forcing us to complement the ‘local’ molecule-based research with integrated approaches that address the properties of the cell as a whole.

#### microRNA

A class of small, non-coding RNAs that are important for development and cell homeostasis, with possible roles in several human disease pathologies.

It is now clearly understood that most cellular functions are carried out by groups of molecules within functional modules<sup>1</sup>. These modules are not isolated from each other; they interact and frequently overlap (for example, see REF. 104), within a network with an inherent scale-free hierarchy, in which the achievable dynamical range is constrained by the underlying topology<sup>53,91</sup>. This organizational framework is shaped during evolution at many levels. The accumulation of local changes that affect the small, highly integrated modules slowly impacts the larger, less integrated modules, which indicates that evolution and natural selection reuse existing modules to further increase the organism's survival probability and its complexity.

This developing framework will significantly alter our understanding of biology and, eventually, will

have important implications for the practice of medicine. The breathtaking advances of modern molecular reductionist biology are starting to pay clinical dividends, from the diagnosis of selected leukaemias, on a molecular level, to their molecularly targeted treatment with, for example, receptor tyrosine kinase inhibitors. Network biology offers the possibility of simultaneous advances in the coming decades. The widespread use of microarrays to refine pathology diagnosis is already evident (for example, see REF. 105). What is lacking is a well-developed framework in which such clinical data can be used to identify modules that are pathologically altered in a given disease state<sup>31</sup>. Once such a framework is developed, the targeted pharmaceutical modification (such as rewiring) of diseased modules will surely follow.

1. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).  
**This influential concept paper strongly argues for the modular organization of biological functions.**
2. Hasty, J., McMillen, D. & Collins, J. J. Engineered gene circuits. *Nature* **420**, 224–230 (2002).
3. Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2002).
4. Koonin, E. V., Wolf, Y. I. & Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420**, 218–223 (2002).
5. Oltvai, Z. N. & Barabási, A.-L. Life's complexity pyramid. *Science* **298**, 763–764 (2002).
6. Wall, M. E., Hlavacek, W. S. & Savageau, M. A. Design of gene circuits: lessons from bacteria. *Nature Rev. Genet.* **5**, 34–42 (2004).
7. Bray, D. Molecular networks: the top-down view. *Science* **301**, 1864–1865 (2003).
8. Alon, U. Biological networks: the tinkerer as an engineer. *Science* **301**, 1866–1867 (2003).
9. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
10. Dorogovtsev, S. N. & Mendes, J. F. F. *Evolution of Networks: from Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, 2003).
11. Bornholdt, S. & Schuster, H. G. *Handbook of Graphs and Networks: from the Genome to the Internet* (Wiley-VCH, Berlin, Germany, 2003).
12. Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
13. Bollobas, B. *Random Graphs* (Academic Press, London, 1998).
14. Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61 (1960).
15. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).  
**This paper introduced the concept of scale-free networks and proposed a mechanism for their emergence.**
16. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
17. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* **268**, 1803–1810 (2001).  
**References 16 and 17 provide the first report on the large-scale organization of metabolic networks, showing its scale-free nature.**
18. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
19. Wagner, A. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283–1292 (2001).
20. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
21. Li, S. *et al.* A map of the interactome network of the metazoan, *C. elegans*. *Science* **2 Jan 2004** (doi:10.1126/science.1091403)
22. Yook, S.-H., Oltvai, Z. N. & Barabási, A.-L. Functional and topological characterization of protein interaction networks. *Proteomics* (in the press).
23. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
24. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574 (2001).
25. Featherstone, D. E. & Broadie, K. Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioessays* **24**, 267–274 (2002).
26. Agrawal, H. Extreme self-organization in networks constructed from gene expression data. *Phys. Rev. Lett.* **89**, 268702 (2002).
27. Wuchty, S. Scale-free behavior in protein domain networks. *Mol. Biol. Evol.* **18**, 1694–1702 (2001).
28. Apic, G., Gough, J. & Teichmann, S. A. An insight into domain combinations. *Bioinformatics* **17**, S83–S89 (2001).
29. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.* **31**, 64–68 (2002).
30. Milo, R., Shen-Orr, S. S., Itzkovitz, S., Kashtan, N. & Alon, U. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).  
**References 29 and 30 introduce the concept of motifs in biological and non-biological networks.**
31. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307–310 (2000).
32. Milgram, S. The small world problem. *Psychol. Today* **2**, 60 (1967).
33. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
34. Chung, F. & Lu, L. The average distances in random graphs with given expected degrees. *Proc. Natl Acad. Sci. USA* **99**, 15879–15882 (2002).
35. Cohen, R. & Havlin, S. Scale-free networks are ultra small. *Phys. Rev. Lett.* **90**, 058701 (2003).
36. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).  
**This paper reports that in protein interaction networks the highly connected nodes tend to link to less connected proteins, which is the so-called disassortative property.**
37. Pastor-Satorras, R., Vázquez, A. & Vespignani, A. Dynamical and correlation properties of the Internet. *Phys. Rev. Lett.* **87**, 258701 (2001).
38. Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
39. Rzhetsky, A. & Gomez, S. M. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* **17**, 988–996 (2001).
40. Qian, J., Luscombe, N. M. & Gerstein, M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313**, 673–681 (2001).
41. Bhan, A., Galas, D. J. & Dewey, T. G. A duplication growth model of gene expression networks. *Bioinformatics* **18**, 1486–1493 (2002).
42. Pastor-Satorras, R., Smith, E. & Sole, R. Evolving protein interaction networks through gene duplication. *J. Theor. Biol.* **222**, 199–210 (2003).
43. Vazquez, A.-F., Flammini, A., Maritan, A. & Vespignani, A. Modeling of protein interaction networks. *ComPlexUs* **1**, 38–44 (2003).
44. Kim, J., Krapivsky, P. L., Kahng, B. & Redner, S. Infinite-order percolation and giant fluctuations in a protein interaction network. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **66**, 055101 (2002).
45. Wagner, A. How the global structure of protein interaction networks evolves. *Proc. R. Soc. Lond. B* **270**, 457–466 (2003).
46. Eisenberg, E. & Levanon, E. Y. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **91**, 138701 (2003).
47. Ravasz, E. & Barabási, A.-L. Hierarchical organization in complex networks. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **67**, 026112 (2003).
48. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).
49. Simon, I. *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697–708 (2001).
50. Tyson, J. J., Cskasz-Nagy, A. & Novak, B. The dynamics of cell cycle regulation. *Bioessays* **24**, 1095–1109 (2002).
51. McAdams, H. H. & Shapiro, L. A bacterial cell-cycle regulatory network operating in time and space. *Science* **301**, 1874–1877 (2003).
52. Bhalla, U. S., Ram, P. T. & Iyengar, R. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science* **297**, 1018–1023 (2002).
53. 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).  
**This paper introduced the concept of hierarchical networks, specifically in the context of metabolism.**
54. Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G. & Alon, U. Subgraphs in random networks. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **68**, 026127 (2003).
55. Wuchty, S., Oltvai, Z. N. & Barabási, A.-L. Evolutionary conservation of motif constituents within the yeast protein interaction network. *Nature Genet.* **35**, 176–179 (2003).
56. Conant, G. C. & Wagner, A. Convergent evolution of gene circuits. *Nature Genet.* **34**, 264–246 (2003).
57. Hinman, V. F., Nguyen, A. T., Cameron, R. A. & Davidson, E. H. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc. Natl Acad. Sci. USA* **100**, 13356–13361 (2003).
58. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. Pseudofractal scale-free web. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* **65**, 066122 (2002).
59. Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. & Dandekar, T. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae*. *Bioinformatics* **18**, 351–361 (2002).
60. Snel, B., Bork, P. & Huynen, M. A. The identification of functional modules from the genomic association of genes. *Proc. Natl Acad. Sci. USA* **99**, 5890–5895 (2002).
61. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99**, 7821–7826 (2002).
62. Holme, P., Huss, M. & Jeong, H. Subnetwork hierarchies of biochemical pathways. *Bioinformatics* **19**, 532–538 (2003).
63. Rives, A. W. & Galitski, T. Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA* **100**, 1128–1133 (2003).

64. Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA* **100**, 12123–12128 (2003).
65. Ihmels, J. *et al.* Revealing modular organization in the yeast transcriptional network. *Nature Genet.* **31**, 370–377 (2002).
66. Bader, G. D. & Hogue, C. W. Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnol.* **20**, 991–997 (2002).
67. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
68. Tomrow, S. & Mewes, H. W. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res.* **31**, 6283–6289 (2003).
69. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
70. Bar-Joseph, Z. *et al.* Computational discovery of gene modules and regulatory networks. *Nature Biotechnol.* **21**, 1337–1342 (2003).
71. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000). **This paper addresses the topological robustness and vulnerability of complex networks.**
72. Winzler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
73. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
74. Gerdes, S. Y. *et al.* Experimental determination and system-level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684 (2003).
75. Yu, B. J. *et al.* Minimization of the *Escherichia coli* genome using a Tn5-targeted Cre/loxP excision system. *Nature Biotechnol.* **20**, 1018–1023 (2002).
76. Kolisnychenko, V. *et al.* Engineering a reduced *Escherichia coli* genome. *Genome Res.* **12**, 640–647 (2002).
77. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. Evolutionary rate in the protein interaction network. *Science* **296**, 750–752 (2002).
78. Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* **13**, 2229–2235 (2003).
79. Dezso, Z., Oltvai, Z. N. & Barabási, A.-L. Bioinformatics analysis of experimentally determined protein complexes in the yeast, *Saccharomyces cerevisiae*. *Genome Res.* **13**, 2450–2454 (2003).
80. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
81. Alon, U., Surette, M. G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168–171 (1999). **References 80 and 81 represents the first theoretical/experimental study on the functional robustness of a cellular sub-network, focusing on the bacterial chemotaxis receptor module.**
82. von Dassow, G., Meir, E., Munro, E. M. & Odell, G. M. The segment polarity network is a robust developmental module. *Nature* **406**, 188–192 (2000).
83. Albert, R. & Othmer, H. G. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J. Theor. Biol.* **223**, 1–18 (2003).
84. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl Acad. Sci. USA* **95**, 8420–8427 (1998).
85. Savageau, M. *Biochemical Systems Analysis: a Study of Function and Design in Molecular Biology* (Addison-Wesley, Reading, 1976).
86. Fell, D. A. *Understanding the Control of Metabolism* (Portland, London, 1997).
87. Schilling, C. H. & Palsson, B. O. The underlying pathway structure of biochemical reaction networks. *Proc. Natl Acad. Sci. USA* **95**, 4193–4198 (1998).
88. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA* **99**, 15112–15117 (2002).
89. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnol.* **19**, 125–130 (2001).
90. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* **420**, 186–189 (2002). **After their theoretical work on flux-balance analysis, the authors of references 89 and 90 show its relevance to predicting experimentally observable metabolic flux values.**
91. Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N. & Barabási, A.-L. Global organization of metabolic fluxes in *E. coli*. *Nature*, (in the press).
92. de la Fuente, A., Brazhnik, P. & Mendes, P. Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet.* **18**, 395–398 (2002).
93. Kuznetsov, V. A., Knott, G. D. & Bonner, R. F. General statistics of stochastic processes of gene expression in eucaryotic cells. *Genetics* **161**, 1321–1332 (2002).
94. Farkas, I. J., Jeong, H., Vicsek, T., Barabási, A.-L. & Oltvai, Z. N. The topology of the transcription regulatory network in the yeast, *Saccharomyces cerevisiae*. *Physica A* **318**, 601–612 (2003).
95. Grigoriev, A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **29**, 3513–3519 (2001).
96. Ge, H., Liu, Z., Church, G. M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **29**, 482–486 (2001).
97. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**, 37–46 (2002).
98. Goh, K. I., Kahng, B. & Kim, D. Fluctuation-driven dynamics of the internet topology. *Phys. Rev. Lett.* **88**, 108701 (2002).
99. Braunstein, L. A., Buldyrev, S. V., Cohen, R., Havlin, S. & Stanley, H. E. Optimal paths in disordered complex networks. *Phys. Rev. Lett.* **91**, 168701 (2003).
100. Menezes, M. A. & Barabási, A.-L. Fluctuations in network dynamics. *Phys. Rev. Lett.* (in the press).
101. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
102. Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395–404 (2003).
103. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
104. Danial, N. N. *et al.* BAD and glucokinase reside in a mitochondrial complex that integrates glycolysis and apoptosis. *Nature* **424**, 952–956 (2003).
105. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
106. Emmerling, M. *et al.* Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. *J. Bacteriol.* **184**, 152–164 (2002).

## Acknowledgements

We thank two anonymous reviewers for their comments and M. Vidal for sharing unpublished work. This research was supported by grants from the National Institutes of Health, Department of Energy (to A.-L.B. and Z.N.O.) and the National Science Foundation (to A.-L.B.).

## Competing interests statement

The authors declare that they have no competing financial interests.

 Online links

## FURTHER INFORMATION

Albert-László Barabási's laboratory: <http://www.nd.edu/~alb>  
Zoltán N. Oltvai's laboratory: <http://www.oltvailab.northwestern.edu>

Self-organized networks: [www.nd.edu/~networks](http://www.nd.edu/~networks)

Access to this interactive links box is free online.