The Journal of Experimental Biology 210, 1548-1558
Published by The Company of Biologists 2007
doi:10.1242/jeb.003731

Biological impacts and context of network theory

Eivind Almaas

Microbial Systems Biology, Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, 7000 East Avenue, PO Box 808, L-452, Livermore, CA 94550, USA

e-mail: almaas@llnl.gov

Accepted 27 March 2007

Summary

Many complex systems can be represented and analyzed as networks, and examples that have benefited from this approach span the natural sciences. For instance, we now know that systems as disparate as the World Wide Web, the Internet, scientific collaborations, food webs, protein interactions and metabolism all have common features in their organization, the most salient of which are their scale-free connectivity distributions and their small-world behavior. The recent availability of large-scale datasets that span the proteome or metabolome of an organism have made it possible to elucidate some of the organizational principles and rules that govern their function, robustness and evolution. We expect that combining the currently separate layers of information from gene regulatory networks, signal transduction networks, protein interaction networks and metabolic networks will dramatically enhance our understanding of cellular function and dynamics.

Glossary available online at
http://jeb.biologists.org/cgi/content/full/210/9/1548/DC1

Key words: systems biology, complex networks, computational biology, protein interaction networks, metabolic networks, optimization.

Introduction

The post-genomic era has brought unprecedented opportunities to bridge and bring together traditionally separate disciplines in the natural sciences. The development of high-throughput techniques and the wide availability of large biological datasets, ranging from annotated genomes to organism-level maps of protein interactions and cellular metabolism, have made it possible simultaneously to probe cellular function at multiple levels. Most of the dramatic progress in the natural sciences during the last century can be directly related to the reductionist approach; the behavior of a system can be predicted and understood solely from the detailed knowledge of the system's elementary constituents.

However, it is by now clear that our ability to understand simple fundamental laws governing the individual building blocks is a far cry from being able to predict the overall behavior of a complex system (Anderson, 1972). Since evolutionary forces have shaped the complex and highly non-linear interactions between genes, proteins and metabolites, there exists considerable variation in the nature of both the elementary building blocks and their interactions, requiring the development of novel methods capable of uncovering cellular organization and functional principles at the systems level.

In this review, I aim to show how computational systems biology (Kitano, 2002), and more particularly network theory as applied to biological systems, offers quantifiable tools to uncover organizational principles of biological systems at the cellular level.

Network analysis of protein interaction systems

In building a network from physical protein binding data, it is customary to consider individual proteins as the nodes, and the existence of a physical interaction between a pair of proteins, e.g. as measured by high-throughput experiments, as a link between two corresponding nodes. Fig. 1A shows the protein interaction network (PIN) for the yeast *C. elegans* using data from various high-throughput experiments available from The BioGRID (version 2.0.20; http://www.thebiogrid.org/). The lowest connectivity nodes (only a single neighbor) are colored blue, nodes with an intermediate connectivity (two to nine) are green, while the highly connected nodes ($\geqslant$10 neighbors) are colored red. This figure suggests that the network is somewhat organized in a layer structure, with the majority of the singly connected nodes at the periphery and the highly connected nodes in the center. However, we need to

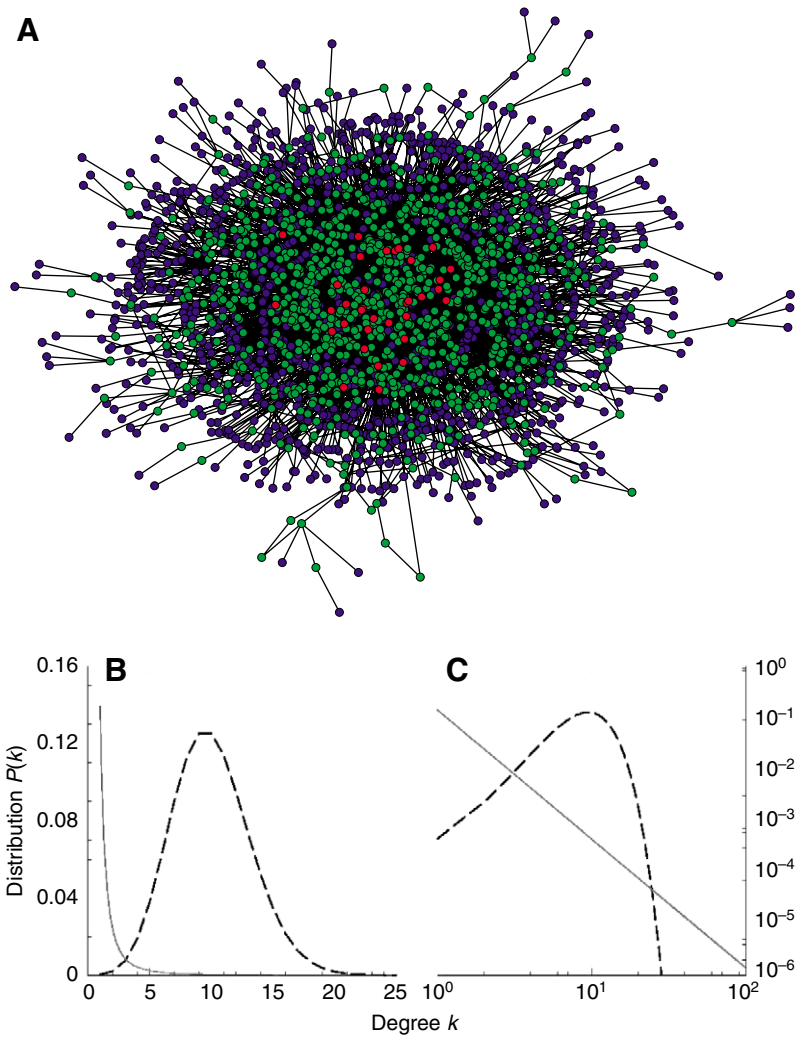

Fig. 1. (A) Protein interaction network of the nematode *Caenorhabditis elegans* using data from The BioGRID (version 2.0.20; http://www.thebiogrid.org/). Nodes (proteins) in blue have a connectivity of one. Nodes in green have a connectivity between two and nine, while the red nodes have a connectivity of ⩾10. Comparison of a linear (B) and logarithmic plot (C) of a Poisson connectivity distribution (broken line) with mean λ=10 and a power-law connectivity distribution (solid line) with exponent γ=2.5.

introduce quantitative, statistical measures to systematically probe the properties of this PIN.

In this example, links represented experimentally measured binding. However, links may represent more general relationships between proteins than just physical binding. For instance, correlations between mRNA expression profiles in microarray data can be used as a basis for the determination of a direct link between two nodes. In this situation, one may define interactions between proteins whose mRNA expression profiles have a correlation value above an appropriately chosen cut-off, say κ, while no links are introduced when the pair-wise correlation values are less than κ.

With the availability of large-scale experimental data on PINs, such as those for *Saccharomyces cerevisiae* (Uetz et al., 2000; Ito et al., 2001; Gavin et al., 2002; Ho et al., 2002) and *Drosophila melanogaster* (Giot et al., 2003), network approaches have become crucial for developing a comprehensive understanding at the organism level. There exist many methods to dissect and analyze networks in general, and the PIN in particular. In the following, I will discuss the most common of these methods, while highlighting the biologically relevant information that can be gleaned.

Applying the tools of network analysis, a system's interacting elements (e.g. genes, proteins or metabolites) are represented as nodes, and the existence of an interaction between two elements as a link between the respective nodes. Links may carry information about the interaction, either as a link weight (interaction strength) or by specifying an interaction asymmetry (link direction). In general, a network consists of N nodes and M links and is represented mathematically as a binary matrix frequently called the 'adjacency matrix' [a_{ij}]. An interaction between the nodes i and j is present when the matrix element $a_{ij}=1$ and absent if $a_{ij}=0$.

Connectivity distribution

The modeling and analysis of systems as disparate as the World Wide Web and PINs has revealed surprising similarities in their structural organization. Possibly the simplest measure to characterize the role that a node (in this section, a protein) plays in the network is the 'node connectivity', or degree, $k_i=\Sigma_j a_{ij}$. We can also define the 'average node degree' in the network, corresponding to an average protein's number of interaction partners, $<k>=(1/N)\Sigma_i k_i$. However, these measures do not provide a detailed insight into the organization of PINs.

To gain a more detailed insight into the structure of the PIN, we study the 'connectivity distribution' given by $P(k)=N_k/N$, where N_k is the number of nodes with k neighbors. From this measure, we may determine the variation in connectivities in the network. Such distributions were studied by Erdös and Rényi (e.g. Bollobás, 2001) and they showed that simple random graphs lead to a Poisson connectivity distribution. However, for many real networks, including the PINs, $P(k)$ does not have a Poisson-type behavior or, even more generally, a unimodal behavior as predicted by the Erdös–Rényi random graph theory. Instead, $P(k)$ is frequently found to adhere to a heavy-tailed distribution that is often modeled as a power-law $P(k)\sim k^{-\gamma}$ (Albert and Barabási, 2002). Fig. 1 shows a side-by-side comparison of a generic Poisson and power-law distribution using linear (Fig. 1B) and logarithmic scales (Fig. 1C). Notably, the logarithmic scale represents the power-law distribution as a straight line, and its decay is clearly seen to be significantly slower than that of the Poisson. Consequently, slowly decaying

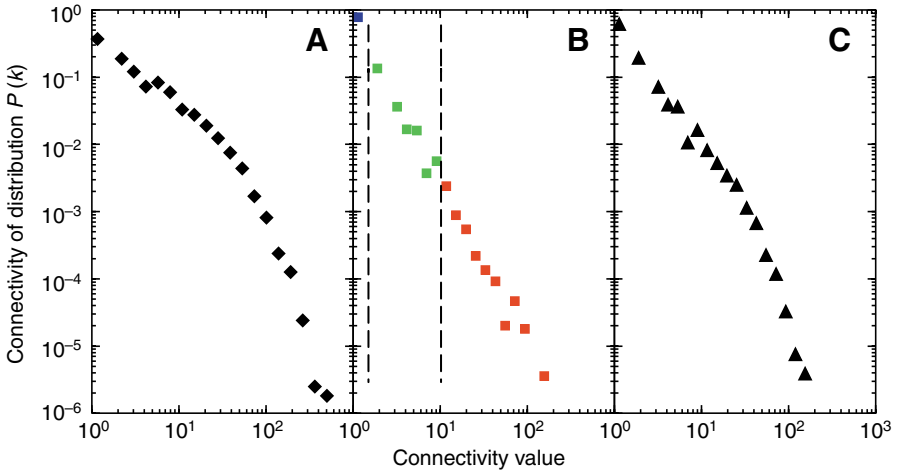

Fig. 2. Connectivity distribution $P(k)$ for the protein interaction networks of (A) the yeast *S. cerevisiae*, (B) the nematode *C. elegans* and (C) the fly *D. melanogaster* from The BioGRID (version 2.0.20; http://www.thebiogrid.org/). The colors in B correspond to the node-colors in Fig. 1; nodes with a connectivity of one are blue, a connectivity between two and nine is green, and highly connected nodes ($\geqslant$10) are red.

distributions such as the power-law are described as being heavy-tailed. Fig. 2 shows the connectivity distribution of the PINs of the yeast *S. cerevisiae*, the nematode *Caenorhabditis elegans* and the fruit fly *D. melanogaster* (see also Table 1).

It is interesting to note that if the connectivity distribution had instead been single-peaked, such as a Poisson or a Gaussian, the notion of a typical node, as described by the average connectivity $<k>$, would be valid. Since the PINs are networks with a heavy-tailed connectivity distribution, the majority of the nodes only have a few interaction partners while they coexist with nodes that participate in hundreds of interactions. Consequently, there exists no typical node in the PINs, and they are frequently described as being 'scale-free'. The class of nodes with a very large number of interaction partners is called a network 'hub'. These hub proteins often have biological properties that are significantly different from those of proteins participating in only a few interactions. Note that no formal definition exists to separate a hub protein from non-hub proteins.

One of the most popular network models that captures the observed heterogeneity of the connectivity distribution was proposed by Barabási and Albert (Barabási and Albert, 1999). It is similar to a model by Price (Price, 1965) [see Newman (Newman, 2003b), for a detailed discussion and comparison of the models]. These models are based on the notion that in a growing and evolving network, new nodes are not connected with uniform probability to already existing nodes. Instead, new nodes have a higher chance of connecting to those with many neighbors than to nodes with few. This is often called the rich-gets-richer effect or 'preferential attachment'. If the chance Π_i of connecting to an already existing node i is linearly proportional to the node degree k_i, i.e. $\Pi_i=k_i/\Sigma_j k_j$, the resulting connectivity distribution is a power-law with an exponent of $\gamma=3$ (Albert and Barabási, 2002; Newman, 2003b).

Note that, if the effective preferential attachment rule is a non-linear function of the degree k, we can no longer expect the resulting connectivity distribution to be scale-free (Krapivsky et al., 2000; Krapivsky and Redner, 2001). In particular, if the preferential attachment rule is slower than linear in k, the connectivity distribution is a stretched exponential. For the case of a preferential attachment rule that is faster than linear in k, the resulting network is of a star type, where the majority of the nodes are connected to a single 'super-hub' (Krapivsky et al., 2000; Krapivsky and Redner, 2001).

Protein interaction networks and evolution

Although the connectivity distribution of PINs is heavy-tailed, which is consistent with the preferential attachment prediction, it is far from clear that the actual evolutionary mechanisms responsible for the current structure of these networks are related to preferential attachment. It appears unlikely that an evolutionary process directly measures the size of a protein's network neighborhood. In fact, multiple alternative processes exist that may give rise to a scale-free connectivity distribution (Newman, 2005). These include local network growth rules, such as gene duplication (addition of nodes) and gene diversification (loss and/or addition of links)

Table 1. *Properties of three whole-organism protein interaction networks available from The BioGRID (version 2.0.20; http://www.thebiogrid.org/)*

Organism	N	$<k>$	S	$<C>$	$<C_{rand}>$	ρ
S. cerevisiae	5298	19.04	5294	0.154	0.0036	–0.040
C. elegans	2774	3.14	2551	0.020	0.0011	–0.159
D. melanogaster	7490	6.67	7372	0.030	0.0089	–0.039

For each network, we have indicated size (N), average node connectivity ($<k>$), size of the giant component (S), average clustering ($<C>$), average clustering for a comparable Erdös–Rényi random network ($<C_{rand}>$) and assortativity (ρ).

(for a review, see Albert and Barabási, 2002), all giving rise to scale-free connectivity distributions. Consequently, models based on local growth mechanisms demonstrate that there are many possible network expansion rules that have an effective linear preferential attachment as a result. Nevertheless, it is possible to directly estimate the evolutionary rates of link addition or removal, as well as those of node duplication from empirical data (Wagner, 2001). Focusing on the yeast PIN, two empirical studies (Eisenberg and Levanon, 2003; Wagner, 2003) clearly support the hypothesis that local network-growth rules give rise to linear preferential attachment, where highly connected proteins display an elevated rate of interaction turnover.

Network clustering

It has long been argued that biological systems are 'functionally modular' (e.g. Hartwell et al., 1999), and it has been a much sought-after goal to understand how this modularity is reflected in the structure of the networks. The 'clustering coefficient' (c) of a node (Watts and Strogatz, 1998):

$$c_i = \frac{1}{k_i(k_i-1)} \sum_{j,l} a_{ij}a_{il}a_{jl} ,$$

measures the degree to which the neighborhood of a node resembles a complete subgraph built from triangles and is the ratio of the actual number of triangles to all possible triangles, for which node i is a member. Consequently, c_i is a measure of the cliquishness, or transitivity, of the local neighborhood. Take Fig. 4C as a network example. Nodes D–B–E are connected in a triangle, while nodes A–C–B–D are connected in a square. The clustering of node A is c_A=0, since there is no direct link between its nearest neighbors, nodes C and D. However, the clustering of node E is c_E=1, since its two (only) neighbors are connected. Finally, the clustering of node B is c_B=1/6, since some of its neighbors (namely nodes D and E) are directly connected.

The average clustering coefficient $<C>=(1/N)\Sigma_i c_i$ provides information on the global distribution of links. A value of $<C>$ close to unity indicates a high level of modularity, or cohesiveness of triangles, in the network, while a value close to zero indicates a lack of modularity. It is customary to test the significance of a particular $<C>$ value by comparison with a suitable random-network model consisting of the same number of nodes and links (Albert and Barabási, 2002). For most such null models, we would find a reference clustering of $<C>_{rand}=<k>/N$, where $<k>=2M/N$ is the average node degree.

Assuming that a network has a non-zero $<C>$, we may further investigate the network's large-scale modularity structure by studying the average clustering as a function of node degree, $C(k)$ (Dorogovtsev et al., 2002). If the network shows a hierarchical modularity (Ravasz et al., 2002), then the clustering $C(k)\sim 1/k$. In this case, nodes with few neighbors tend to have network neighborhoods with high clustering, while the highly connected nodes act as bridges tying different parts of the network together. However, the network modules are not clearly discernible, being interwoven on all levels.

Network assortativity

In many real networks, there exist correlations between the properties of neighboring nodes. In particular, it is often the case that the connectivity of neighboring nodes is correlated. When these correlations are absent, we can expect that the joint probability of two randomly selected nodes i and j having k_i and k_j neighbors, respectively, is $P(k_i,k_j)=P(k_i)P(k_j)$. However, in the presence of such node–node correlations, knowing the connectivity k_i of node i, we have received information about the connectivity of any node j directly connected to node i with a link. Several methods have been developed to measure these connectivity correlations, and we will highlight two of them (Maslov and Sneppen, 2002; Pastor-Satorras et al., 2001; Newman, 2002; Newman, 2003a).

The first method to measure correlations between neighboring nodes was suggested by Vespignani and co-workers (Pastor-Satorras et al., 2001). It measures connectivity correlations by calculating the 'neighborhood connectivity' of a node $k_{nn,i}=(1/k_i)\Sigma_j k_j a_{ij}$, where index nn denotes 'nearest neighbor'. Consequently, $k_{nn,i}$ measures the affinity with which a node i connects to other nodes of either high or low degrees. In Fig. 3, we have plotted the function $k_{nn}(k)$, which is the average neighborhood degree for nodes with connectivity k. Note that if $k_{nn}(k)$ is an increasing function of k, the network shows an 'assortative' mixing, and high-degree nodes preferentially tend to be connected to other high-degree nodes. For the opposite situation, where $k_{nn}(k)$ is a decreasing function of k (as in Fig. 3B), low-degree nodes tend to be connected to high-degree nodes, and the network is 'disassortative'. This is also the typical case for computer networks, where a limited number of servers each are connected to a large number of individual computers (Pastor-Satorras et al., 2001).

The second method of measuring degree–degree correlations in a network is the Pearson correlation, ρ, in nearest neighbor degrees, called the 'assortativity' (Newman, 2002). A Pearson correlation is often interpreted as a measure of a linear relationship between two variables, in this case the connectivity of node pairs joined by a link. The degree–degree correlation ranges from ρ=1 to ρ=–1. The distribution $k_{nn}(k)$ and the assortativity ρ are related as follows. If $k_{nn}(k)$ is uniform, then ρ=0. However, if $k_{nn}(k)$ is increasing or decreasing, then ρ is positive or negative, respectively. The magnitude of ρ indicates the strength of the correlation. It is straightforward to develop similar expressions for directed networks (Newman, 2003a).

The last column of Table 1 shows the assortativity ρ for three whole-organism-level PINs. As expected, the trends displayed in Fig. 3 agree with the assortativity correlations calculated using ρ (Newman, 2002). In particular, Fig. 3A and Fig. 3C show no clear increasing or decreasing trend in $k_{nn}(k)$, which agrees well with the calculated assortativity values being close to zero. Taken together, these two methods offer detailed insights into the connectivity correlations of a network.

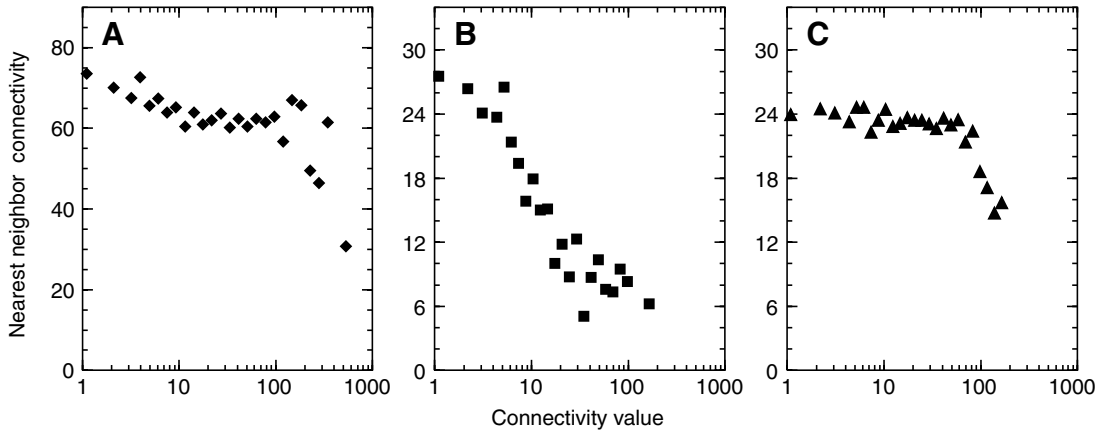

Fig. 3. Average nearest neighbor connectivity $k_{nn}(k)$ for the protein interaction networks of (A) *S. cerevisiae*, (B) *C. elegans* and (C) *D. melanogaster* from The BioGRID (version 2.0.20; http://www.thebiogrid.org/).

Protein interaction networks and essentiality

So far we have discussed topological properties of PINs without emphasizing the close relationship between network representations and biological information. The first indication that the large-scale structure of a PIN network might carry biological information arose from investigations of network robustness (Albert et al., 2000). This work demonstrated that networks with heavy-tailed connectivity distributions were robust against random failures, yet fragile when an attack occurred at a highly connected node. The robustness of a network was evaluated in terms of a network topology measure, the 'giant component'. A connected component consists of all nodes between which there exists a path, and the giant component is the largest among the connected components. The third column of Table 1 lists the giant component of the three PINs. We can study the resilience of a network to node removal by monitoring the size S of the giant component while randomly deleting nodes from the network (corresponding to failure) or iteratively removing the largest hubs (corresponding to attack). Through such a node-removal analysis, it was discovered that networks with a scale-free connectivity distribution retain a giant component while subject to random failures (Albert et al., 2000). On the other hand, when the scale-free networks are subject to attack, they fragment very quickly. Consequently, these networks are extremely robust against random perturbations, yet highly susceptible to targeted attacks.

Several molecular biology techniques are now available for the experimental perturbation and disruption of PINs. In fact, a large-scale experimental study in *S. cerevisiae* shows that only 18.7% of the total number of genes are essential on disruption or removal (Giaever et al., 2002), while a study on *Escherichia coli* found 13.7% of the genes to be essential (Gerdes et al., 2003). Motivated by the above theoretical and experimental observations on network fragility, Barabási and co-workers investigated the possibility of correlations between a protein's connectivity and its phenotypic essentiality, discovering an increasing likelihood for highly connected proteins to be essential (Jeong et al., 2001). In other words, the more interaction partners a protein has, the more likely it is to be involved in an essential cellular function. This result is often called the 'centrality-lethality' rule. Although recently debated (Coulomb et al., 2005), careful analyses strongly support the centrality-lethality rule (Batada et al., 2006).

A recent study suggests that the increased lethality of highly connected proteins can be explained by a simple mechanism (He and Zhang, 2006). The idea is to explain the centrality-lethality rule by assuming that essential nodes and 'links' are randomly distributed on the network. The function of an essential link is carried out by the interaction of the incident proteins, and both nodes are essential. This model generates the centrality-lethality rule through the simple fact that it is more likely for a hub to partake in an essential link than for a low-degree node. By choosing the essential link and node fractions appropriately, it is possible to fit the observed centrality-lethality rule within experimental error bars (He and Zhang, 2006).

Since highly connected proteins occupy a special role in the network, it is interesting to study if hub proteins should evolve at a different pace from proteins with only a few interaction partners (Batada et al., 2006). Indeed, because highly connected proteins do not have a higher density of active domains, they do not show any significant difference in mean rate of protein evolution. However, the hub proteins of *S. cerevisiae* contain a higher number of phosphorylation sites than do non-hub proteins and show a marked trend of being encoded by mRNAs with short half-lives. This indicates that highly connected proteins are subject to much tighter control, being part of dynamic short-lived protein complexes (Batada et al., 2006).

Protein interaction networks and dynamics

We have focused on the static aspects of a PIN, but proteins are constantly being degraded and produced and many carry out their functions in specific cellular locations such as a cellular membrane. A more realistic depiction would address

the temporal and spatial aspects of the situation. Whole-organism protein-expression arrays are currently unavailable, and the chosen substitute has been the mRNA expression array. A recent analysis (Han et al., 2004) indicates that highly connected nodes in the *S. cerevisiae* PIN are either 'date-hubs', binding to their partners at different times or locations, or 'party-hubs', which interact with most of their network neighbors simultaneously. Including temporal aspects in the PIN analysis allows for the investigation of information flow, since the temporal activation of protein transcription is reflective of evolved regulatory mechanisms to ensure proper cellular responses to external stimuli.

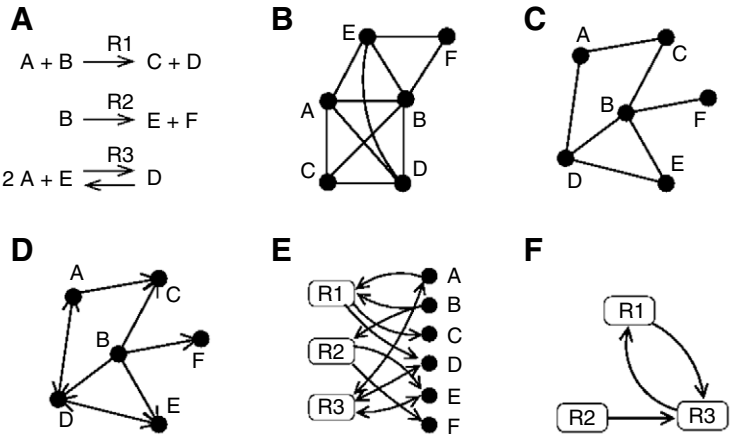

Fig. 4. Cellular metabolism can be represented as a network. (A) Toy metabolic reaction set. Network description of the reaction set: (B) connecting all metabolites in a single reaction with undirected links; (C) substrates are only connected to products with undirected links; and (D) same as in C with directed links. (E) Bipartite network representation of the reaction set. (F) Network with reactions as nodes, and reactions that share a metabolite as educt–product are connected.

Network analysis of metabolism

Cellular metabolism depends on enzymatic reactions where substrates, such as glucose or acetate, are converted into products by enzymes. However, the set of metabolic reactions can be translated into a network representation in many different ways. Fig. 4 demonstrates several possible network representations of a simple metabolic reaction set. Fig. 4A describes the relationship between the metabolites A–F. In the first reaction, A+B→C+D, we say that A and B are educts and C and D are products. A common network representation is displayed in Fig. 4C, where metabolites are nodes, and two metabolites are connected with an undirected link if they participate as an educt and a product, respectively, in the same reaction. Note that a link does not represent a single reaction, or enzyme, as two metabolites may appear in multiple reactions. An example of this possibility is shown in Fig. 4A, where metabolites A and D co-occur in reactions R_1 and R_3, and the link between A and D in Fig. 4C corresponds to both reactions. To further complicate the mapping, one reaction may also appear as multiple links (see Fig. 4). An alternative representation is that of a bipartite network (Fig. 4E), where the two kinds of nodes are metabolites or enzymes. For this case, a directed link from (to) a metabolite to (from) an enzyme indicates that the metabolite acts as an educt (product) in that reaction. Finally, a metabolic reaction set may also be represented as a reaction–reaction network (Fig. 4F). Here, the nodes are reactions and a (possibly directed) link is included between two nodes (reactions) *i* and *j* if a metabolite is used as an educt (product) in reaction *i* and as a product (educt) in reaction *j*.

Metabolic network structure

The various network representations of Fig. 4 have different statistical properties. Using the bacterial metabolism in *E. coli* as an example, Fig. 5 shows the differences in the connectivity distribution, $P(k)$, implied by the three network representations detailed in Fig. 4B–D. Note that $P(k)$ is heavy-tailed in all panels of Fig. 5; however, the result is not as simple for a bipartite network representation (Fig. 4E). In this case, it is possible to distinguish between metabolites and enzymes; for the metabolites, the connectivity distribution is heavy tailed, while the enzyme distribution is best fit by an exponential. This is not surprising, as cofactors such as ATP or NADP may contribute to hundreds of reactions while an enzyme has a limited number of active domains. To further contrast and compare potential biases of various network representations, Table 2 shows the

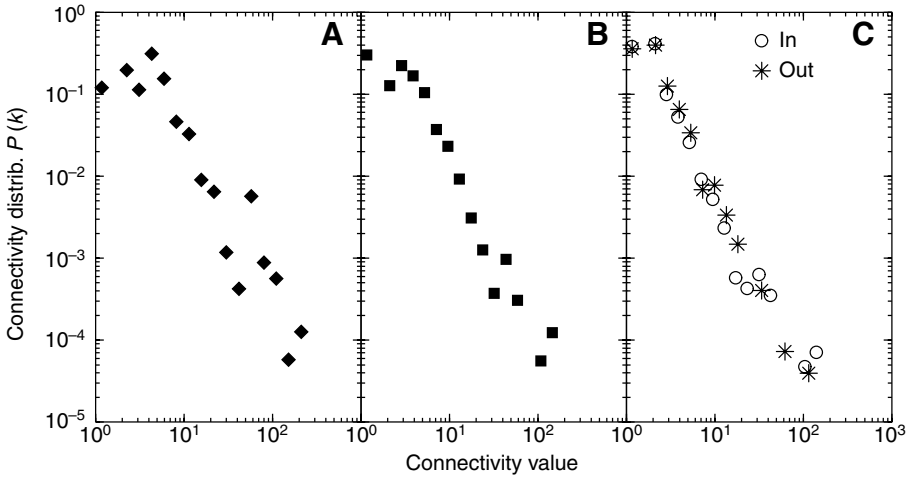

Fig. 5. Connectivity distributions $P(k)$ of *E. coli* metabolism using the three metabolic network representations in Fig. 4. Panel A corresponds to Fig. 4B; B corresponds to Fig. 4C; C corresponds to Fig. 4D.

Table 2. *Average clustering and assortativity for three organismal metabolic networks using the network representations described in Fig. 4B,C*

Organism	N	M_B	M_C	$<C>_B$	$<C>_C$	ρ_B	ρ_C
H. pylori	489	4058	1920	0.72	0.28	–0.285	–0.261
E. coli	540	3753	1867	0.66	0.20	–0.251	–0.217
S. cerevisiae	1064	6941	4031	0.67	0.23	–0.182	–0.150

Abbreviations: N, number of nodes; M, number of links; $<C>$, average clustering; ρ, assortativity; subscript B and C, network representations shown in Fig. 4B and Fig. 4C, respectively.

clustering $<C>$ and the assortativity ρ for three organisms using the network representations of Fig. 4B,C. As expected, the clustering and assortativity corresponding to Fig. 4B is significantly higher than that of Fig. 4C, since the network representation in the former implies a fully connected subgraph for each reaction.

Weighted metabolic networks

The majority of network studies have focused on topological properties and not on the rate of metabolic activity, which can vary significantly from reaction to reaction. This important function is not captured by standard topological approaches. It is necessary to include this information in the network description to develop an understanding of how the structure of a metabolic network affects metabolic activity. A meaningful understanding requires us to consider the intensity (i.e. strength), the direction (when applicable) and the temporal aspects of the interactions. Although much is still unknown about the temporal aspects of metabolic activity inside a cell, recent results have provided information about the relative intensities of the interactions in single-cell metabolism (Sauer et al., 1999; Canonaco et al., 2001; Gombert et al., 2001; Emmerling et al., 2002; Fischer and Sauer, 2003; Cannizzaro et al., 2004; Blank et al., 2005; Fischer and Sauer, 2005). We may incorporate these results into the network analysis by considering links not only to be present or absent, but additionally to carry a 'link weight' that reflects the non-uniform interaction strength between two nodes. A natural, although not unique, measurement of interaction strength for a metabolic network is the amount of substrate being converted to a product per unit time, called the 'flux' of the reaction.

A simple linear optimization approach, called 'flux-balance analysis' (FBA), enables us to calculate the flux rate for each reaction in a whole-cell metabolic network. The FBA method is based on the assumption that the concentration of all cellular metabolites, $[A_i]$, not subject to transport across the cell membrane must satisfy the steady-state constraint of $d[A_i]/dt=\Sigma_j S_{ij}v_j=0$, where S_{ij} is the stoichiometric coefficient of metabolite A_i in reaction j, t is time, and v_j is the steady-state flux of reaction j. We follow the convention that $S_{ij}<0$ ($S_{ij}>0$) if metabolite i is a substrate (product) in reaction j. Take Fig. 4A as an example. The stoichiometric coefficients of reaction $j=R_3$ are then $S_{A,R3}=-2$, $S_{E,R3}=-1$, $S_{D,R3}=1$, while $S_{B,R3}=S_{C,R3}=S_{F,R3}=0$. Note that any flux value v_i satisfying the steady-state constraint corresponds to a stoichiometrically allowed state of the cell. To select flux values that are biologically relevant, we optimize for cellular growth. Experiments support this hypothesis in several conditions, but there are also other meaningful objectives. See Bonarius et al. (Bonarius et al., 1997) and Kauffman et al. (Kauffman et al., 2003) for a more detailed discussion of FBA.

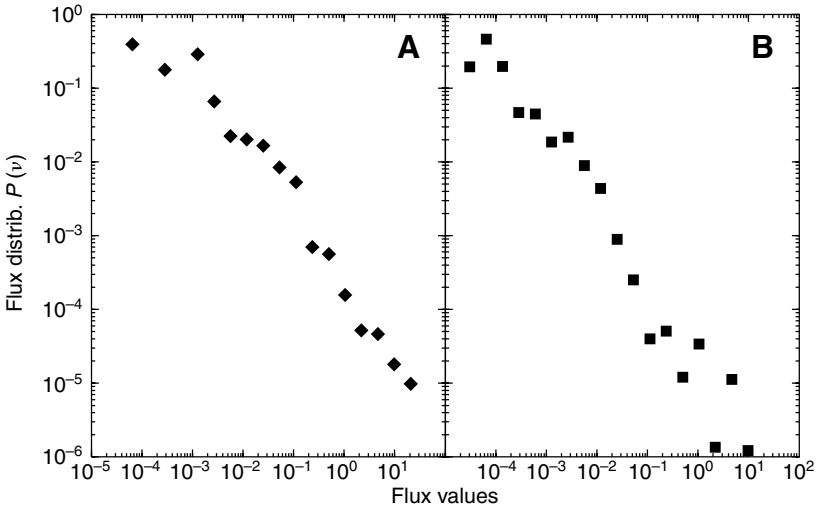

Fig. 6. Distribution of metabolic reaction flux values (link weights) from FBA analysis for the metabolic network of the budding yeast *S. cerevisiae* in (A) aerobic, glucose-limited and (B) aerobic, acetate-limited conditions.

The recent advances in whole-genome annotation has made it possible to generate high-fidelity whole-cell level metabolic networks. Metabolic models of the prokaryotic *Helicobacter pylori* and *E. coli*, as well as the eukaryote *S. cerevisiae*, have been used to predict 'essential genes' (Edwards and Palsson, 2000; Schilling et al., 2002; Duarte et al., 2004; Papp et al., 2004), 'epistatic interactions' where the action of one gene is modified by one or multiple genes at different loci (Segre et al., 2005), and possible 'minimal microbial genomes' (Burgard et al., 2001; Pal et al., 2006). The resulting fluxes from FBA measure each reaction's relative activity. In particular, Almaas et al. demonstrate that, similar to the degree distribution, the flux distribution of *E. coli* displays a strong overall

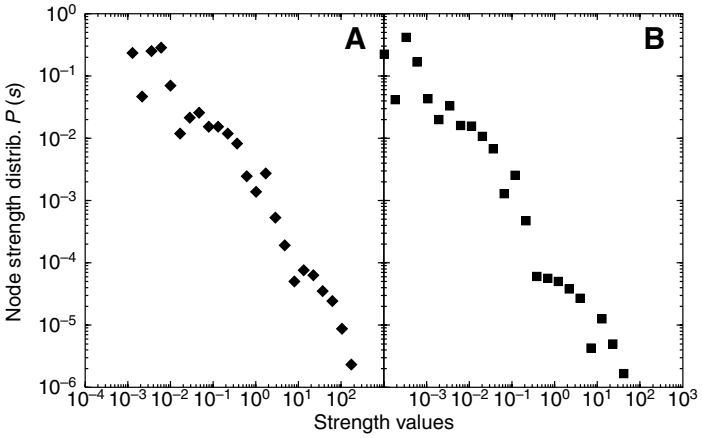

Fig. 7. Distribution of node strength values for *S. cerevisiae* metabolism in (A) aerobic, glucose-limited and (B) aerobic, acetate-limited conditions.

inhomogeneity: reactions with fluxes spanning several orders of magnitude coexist in the same environment (Almaas et al., 2004). Applying the FBA computational approach, the flux distribution for *S. cerevisiae* (Fig. 6) is heavy-tailed, indicating that $P(v)\sim v^{-\alpha}$ with a flux exponent of $\alpha=1.5$. In a recent experiment, the strength of the various fluxes of the central metabolism of *E. coli* was measured using nuclear magnetic resonance (NMR) methods (Emmerling et al., 2002), revealing a power-law flux dependence $P(v)\sim v^{-1}$ (Almaas et al., 2004). This power-law behavior indicates that a vast majority of reactions with small fluxes coexists with a few reactions that have large fluxes.

The FBA approach allows us to analyze the metabolic network as a weighted network since each reaction is assigned a flux value. Such a generalization of non-weighted network measures was originally introduced in the context of the airline transportation and co-authorship networks (Barrat et al., 2004). The first of the generalized network measures is called the 'node strength', s_i, of a node i, defined as $s_i=\Sigma_j w_{ij}a_{ij}$, where w_{ij} is the weight of the link connecting nodes i and j, and a_{ij} is the adjacency matrix as before. The node strength acts as a generalization of the node degree to weighted networks and sums the total weight on the links connected to a node. Fig. 7 shows the distribution of node strengths, $P(s)$, for *E. coli* metabolism with glucose as the single carbon source.

We continue by generalizing the clustering coefficient to weighted networks. Since c_i indicates the local density of triangles, a similar definition with link-weights should make it possible to discern if large or small weights are more or less likely to be found clustered together. We denote one possible definition given by Barrat et al. (Barrat et al., 2004) as $c_{w,i}$, and the average weighted clustering is $<C_w>=(1/N)\Sigma_i c_{w,i}$. If no correlations exist between weights and topology, this new definition of clustering coefficient is equal to that of the unweighted network. Furthermore, we may identify two possible scenarios. If $<C_w>$ is greater than $<C>$, large weights are predominantly distributed in local clusters, whereas if $<C_w>$ is less than $<C>$, triangles are built using mostly low-weight links. Other possible definitions of a weighted clustering coefficient with somewhat different properties have been proposed (Onnela et al., 2005; Zhang and Horvath, 2005; Holme et al., 2007).

Fluxes and metabolic network structure

The flux distributions of a metabolic network rely on the network topology. Some of this dependence is understood by studying the correlation between w_{ij}, the strength of the link connecting nodes i and j and their respective connectivities, k_i and k_j. The metabolic fluxes scale as $<w_{ij}>\sim(k_i k_j)^{\theta}$, where $\theta=0.5$ under glucose-limited conditions in *S. cerevisiae* (Fig. 8A) and *E. coli* (Macdonald et al., 2005), as well as the World-Air-Transportation network (Barrat et al., 2004). We may also find similar behavior in network models. As an example, the betweenness-centrality [a measure of how many shortest paths utilize a given node or link (see Brandes, 2001; Freeman, 1977; Newman, 2001; Wasserman and Faust, 1994) on the Barabási–Albert network model (Fig. 8C)]. However, other values for θ are possible, as demonstrated in Fig. 8B, where we find $\theta=0.7$ for metabolic fluxes under acetate-limited conditions.

How does the network structure influence flux patterns on the level of single metabolites? The observed scale-free flux distribution is compatible with two quite different potential local flux structures. A homogeneous local organization would imply that all reactions producing (consuming) a given metabolite have comparable flux values. On the other hand, a more de-localized, or 'hot backbone', is expected if the local flux organization is heterogeneous, such that each metabolite has a dominant source (consuming) reaction. To distinguish between these two scenarios, we define the measure $Y(k,i)$ (Barthelemy et al., 2003; Almaas et al., 2004) for each metabolite produced or consumed by k reactions, with the following characteristics. If all reactions producing (consuming) metabolite i have comparable values, $Y(k,i)\approx 1/k$. If, however, the activity of a single reaction dominates, then $Y(k,i)\approx 1$, i.e. $Y(k,i)$ is independent of k. For the two cases where the *E. coli* metabolic performance is optimized with glucose and succinate as the only available carbon sources, $Y(k)\sim k^{-0.27}$. This is an intermediate behavior between the two extreme cases described above. However, the exponent value of $\beta=-0.27$ indicates that the large-scale inhomogeneity observed in the overall flux distribution is increasingly valid at the level of the individual metabolites as well.

Consequently, for most metabolites, a single reaction can be identified that dominates its production or consumption. A simple algorithm is capable of extracting the sub-network solely consisting of these dominating reactions, called the 'high-flux backbone' (HFB) (Almaas et al., 2004). This algorithm has the following two steps: (1) for each metabolite, discard all incoming and outgoing links except the two links that dominate mass production; and (2) from the resulting set of reactions, keep only those reactions that appear as both a maximal producer and a maximal consumer.

Note that the resulting HFB is specific to the particular choice of system boundary conditions (i.e. environment). Interestingly, the HFB mostly consists of reactions linked together, forming a giant component with a star-like topology that includes almost all metabolites produced in a specific growth environment. Only a few pathways are disconnected; while these pathways are members of the HFB, their end-products serve only as the second most important source for some other HFB metabolite. One may further analyze the properties of the HFB (Almaas et al., 2004); however, we limit our discussion and simply mention that groups of individual HFB reactions largely agree with the traditional, biochemistry-based partitioning of cellular metabolism into pathways. For example, in the *E. coli* metabolic model, all metabolites of the citric acid cycle are recovered, and so are a considerable fraction of other important pathways, such as those being involved in histidine, murein and purine biosynthesis, to mention a few. While the detailed nature of the HFB depends on the particular growth conditions, the HFB captures the reactions that dominate the metabolic activity for this condition. As such, it offers a complementary approach to elementary flux mode and extreme pathway analyses (Schuster and Hilgetag, 1994; Schilling et al., 2000; Papin et al., 2004), which successfully determine the available modes of operation for smaller metabolic sub-networks.

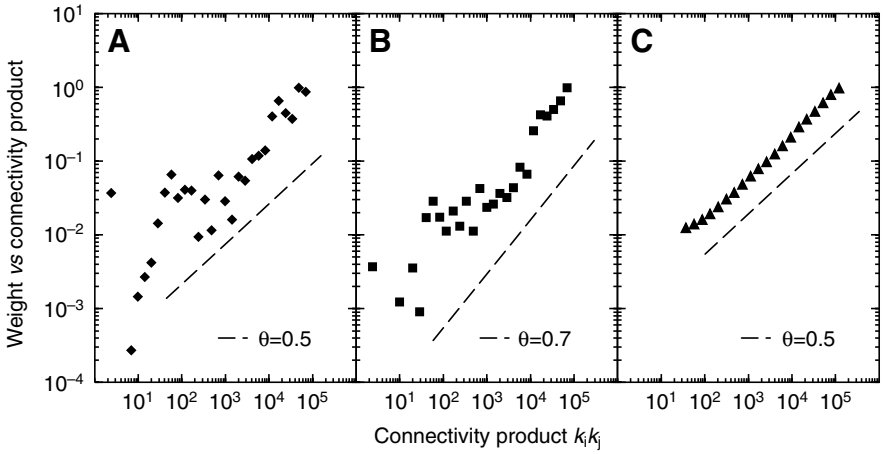

Fig. 8. Correlation between (normalized) link weights and local connectivity for (A) metabolic fluxes in *S. cerevisiae* in glucose-limited and (B) acetate-limited conditions, as well as (C) betweenness-centrality for the Barabási–Albert model. The broken lines serve as visual guides only.

Metabolic core reactions

Any whole-cell metabolic model contains a number of transport reactions for the uptake of nutrients and excretion of byproducts. Consequently, we may systematically sample among all possible environments captured by the model through varying the constraints on uptake reactions. This analysis suggests that optimal metabolic flows are adjusted to environmental changes through two distinct mechanisms (Almaas et al., 2004). The more common mechanism is 'flux plasticity', involving changes in the fluxes of already active reactions when the organism is shifted from one growth condition to another. For example, changing from glucose- to succinate-rich media altered the flux of 264 *E. coli* reactions by more than 20%. Less commonly, environmental changes may induce 'structural plasticity', resulting in changes to the metabolism's active wiring diagram, turning on previously zero-flux reactions and inhibiting previously active pathways. For example, when shifting *E. coli* cells from glucose- to succinate-rich media, 11 previously active reactions were turned off completely, while nine previously inactive reactions were turned on.

The 'metabolic core' is the set of reactions found to be active (carrying a non-zero metabolic flux) in all tested environments. In recent computational experiments where more than 30 000 possible environments were sampled, the metabolic core contained 138 of the 381 metabolic reactions in the model of *H. pylori* (36.2%), 90 of 758 in *E. coli* (11.9%) and 33 of 1172 in *S. cerevisiae* (2.8%) (Almaas et al., 2005). While these reactions respond to environmental changes only through flux-based plasticity, the remaining reactions are conditionally active, being turned on only in specific growth conditions.

The metabolic core can be further partitioned into two types of reactions. The first type consists of those that are essential for biomass formation under all environmental conditions (81 out of 90 reactions in *E. coli*), while the second type of reaction is required only to assure optimal metabolic performance. In case of the inactivation of the second type, alternative sub-optimal pathways can be used to ensure cellular survival. However, the compact core of *S. cerevisiae* only contains reactions predicted by FBA to be indispensable for biomass formation under all growth conditions. A similar selection of metabolic reactions was suggested by Burgard et al. (Burgard et al., 2001). Their 'minimal reaction' contains the metabolic core as well as all reactions necessary for the sustained growth on any chosen substrate. A different definition of a minimal reaction set was proposed by Reed and Palsson (Reed and Palsson, 2004), which consists of the 201 reactions that are always active in *E. coli* for all 136 aerobic and anaerobic single-carbon-source 'minimal environments' capable of sustaining optimal growth.

A reasonable speculation is that the reactions in the metabolic core play an important role in the maintenance of crucial metabolic functions since they are active under all environmental conditions. Consequently, the absence of individual core reactions may lead to significant metabolic disruptions. This hypothesis is strengthened through cross-correlation with genome-scale gene-deletion data (Gerdes et al., 2003): 74.7% of those *E. coli* enzymes that catalyze core metabolic reactions (i.e. core enzymes) are essential, compared with a 19.6% lethality fraction for the non-core enzymes. A similar pattern of elevated essentiality is also

present when analyzing large-scale deletion data for *S. cerevisiae* (Giaever et al., 2002). Here, essential enzymes catalyze 84% of the core reactions, whereas the conditionally active enzymes have an average essentiality of only 15.6% (Almaas et al., 2005). The likelihood that the cores contain such a large concentration of essential enzymes by chance is minuscule, with *P*-values of 3.3×10^{-23} and 9.0×10^{-13} for *E. coli* and yeast, respectively.

Metabolic core reactions also stand apart from the conditionally active ones when comparing their evolutionary conservation. In comparing the core enzymes of *E. coli* with a reference set of 32 bacteria, the average core conservation rate is 71.1% ($P<10^{-6}$) while the non-core enzymes have a homology matching of only 47.7%. Taking into account correlations between essentiality and evolutionary conservation, one would expect the core enzymes to show a conservation level of 63.4% (Almaas et al., 2005).

These results indicate that an organism's ability to adapt to changing environmental conditions rests largely on the continuous activity of the metabolic core, regardless of the environmental conditions, while the conditionally active metabolic reactions represent the different ways in which a cell is capable of utilizing substrates from its environment. This suggests that the core enzymes that are essential for biomass formation, both for optimal and suboptimal growth, may provide effective antibiotic targets, given the cell's need to maintain the activity of these enzymes in all conditions.

Outlook

Network approaches provide an important set of tools to analyze and dissect complex systems spanning from biology to the social sciences. Their generic applicability has successfully been exploited by bringing measures to bear on biological problems that, for example, were originally developed for transportation systems (Albert and Barabási, 2002). As the focus of this review has been PINs and metabolism, a variety of network approaches have given us the opportunity to interrogate the interconnected nature of cellular networks. It is, however, important to remember that the cell is far from a static environment, and it is absolutely necessary to develop new approaches to incorporate both the temporally and spatially dynamic nature of biological systems.

To achieve an accurate description of cellular networks, we also need to couple the available information on gene regulatory, signal transduction, protein interaction and metabolic networks. So far, the majority of research has been focused on studying these networks as separate entities. In particular, the study of metabolism has already shown great promise for coupling to transcription regulatory networks (Covert et al., 2004). Although our current knowledge of kinetic parameters is severely limited, making the development of detailed kinetic models largely intractable, approaches such as FBA married with network methods have opened the door for organism-level investigations of quasi-dynamic cellular response to external and internal perturbation.

The author wishes to thank Profs Holder and Livingstone at Trinity University for helpful discussions and input. This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48, and supported by LDRD Grant 06-ERD-061.

References

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47-97.

Albert, R., Jeong, H. and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* **406**, 378-382.

Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N. and Barabási, A.-L. (2004). Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**, 839-843.

Almaas, E., Oltvai, Z. N. and Barabási, A.-L. (2005). The activity reaction core and plasticity in metabolic networks. *PLoS Comput. Biol.* **1**, e68.

Anderson, P. W. (1972). More is different. *Science* **177**, 393-396.

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

Barrat, A., Barthelemy, M., Pastor-Satorras, R. and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **101**, 3747-3752.

Barthelemy, M., Gondran, B. and Guichard, E. (2003). Spatial structure of the internet traffic. *Physica. A* **319**, 633-642.

Batada, N. N., Hurst, L. D. and Tyers, M. (2006). Evolutionary and physiological importance of hub proteins. *PLoS Comp. Biol.* **2**, 0748.

Blank, L. M., Kuepfer, L. and Sauer, U. (2005). Large-scale c-13-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome. Biol.* **6**, R49.

Bollobás, B. (2001). *Random Graphs.* New York: Academic Press.

Bonarius, H. P. J., Schmid, G. and Tramper, J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends. Biotechnol*. **15**, 308-314.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Soc*. **25**, 163-177.

Burgard, A. P., Vaidyaraman, S. and Maranas, C. D. (2001). Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnol. Progr*. **17**, 791-797.

Cannizzaro, C., Christensen, B., Nielsen, J. and von Stockar, U. (2004). Metabolic network analysis on *Phaffia rhodozyma* yeast using c-13-labeled glucose and gas chromatography-mass spectrometry. *Metab. Eng*. **6**, 340-351.

Canonaco, F., Hess, T. A., Heri, S., Wang, T. T., Szyperski, T. and Sauer, U. (2001). Metabolic flux response to phosphoglucose isomerase knock-out in *Escherichia coli* and impact of overexpression of the soluble transhydrogenase Udha. *FEMS. Microbiol. Lett.* **204**, 247-252.

Coulomb, S., Bauer, M., Bernard, D. and Marsolier-Kergoat, M. C. (2005). Gene essentiality and the topology of protein-interaction networks. *Proc. R. Soc. Lond. B. Biol. Sci.* **272**, 1721-1725.

Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92-96.

Dorogovtsev, S. N., Goltsev, A. V. and Mendes, J. F. F. (2002). Pseudofractal scale-free web. *Phys. Rev. E* **65**, 066122.

Duarte, N. C., Herrgard, M. J. and Palsson, B. O. (2004). Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome. Res*. **14**, 1298-1309.

Edwards, J. S. and Palsson, B. O. (2000). The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA* **97**, 5528-5533.

Eisenberg, E. and Levanon, E. Y. (2003). Preferential attachment in the protein network evolution. *Phys. Rev. Lett*. **91**, 138701.

Emmerling, M., Dauner, M., Ponti, A., Fiaux, J., Hochuli, M., Szyperski, T., Wuthrich, K., Bailey, J. E. and Sauer, U. (2002). Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. *J. Bacteriol*. **184**, 152-164.

Fischer, E. and Sauer, U. (2003). Metabolic flux profiling of *Escherichia coli.* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem*. **270**, 880-891.

Fischer, E. and Sauer, U. (2005). Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat. Genet*. **37**, 636-640.

Freeman, L. C. (1977). A set of measures of centrality based upon betweenness. *Sociometry* **40**, 35-41.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Micho, N. A. M., Cruciat, C. M. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147.

Gerdes, S., Scholle, M., Campbell, J., Balazsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S. et al. (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol*. **185**, 5673-5684.

Giaever, G., Chu, A., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B. et al. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387-391.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E. et al. (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727-1736.

Gombert, A. K., dos Santos, M. M., Christensen, B. and Nielsen, J. (2001). Network identification and flux quantification in the central metabolism of *Saccharomyces cerevisiae* under different conditions of glucose repression. *J. Bacteriol*. **183**, 1441-1445.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P. et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88-93.

Hartwell, L. H., Hopfield, J. J., Leibler, S. and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* **402**, C47-C52.

He, X. and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genet.* **2**, 0826.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K. et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180-183.

Holme, P., Park, S. M., Kim, B. J. and Edling, C. R. (2007). Korean university life in a network perspective: dynamics of a large affiliation network. *Physica. A* **373**, 821-830.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569-4574.

Jeong, H., Mason, S., Barabási, A.-L. and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41-42.

Kauffman, K. J., Prakash, P. and Edwards, J. S. (2003). Advances in flux balance analysis. *Curr. Opin. Biotechnol*. **14**, 491-496.

Kitano, H. (2002). Computational systems biology. *Nature* **420**, 206-210.

Krapivsky, P. L. and Redner, S. (2001). Organization of growing random networks. *Phys. Rev. E* **63**, 066123.

Krapivsky, P. L., Redner, S. and Leyvraz, F. (2000). Connectivity of growing random networks. *Phys. Rev. Lett*. **85**, 4629-4632.

Macdonald, P., Almaas, E. and Barabási, A.-L. (2005). Minimum spanning trees on weighted scale-free networks. *Europhys. Lett*. **72**, 308-314.

Maslov, S. and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* **296**, 910-913.

Newman, M. E. J. (2001). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* **64**, 016132.

Newman, M. E. J. (2002). Assortative mixing in networks. *Phys. Rev. Lett*. **89**, 208701.

Newman, M. E. J. (2003a). Mixing patterns in networks. *Phys. Rev. E* **67**, 026126.

Newman, M. E. J. (2003b). The structure and function of complex networks. *SIAM Rev.* **45**, 167-256.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemp. Phys*. **46**, 323-351.

Onnela, J.-P., Saramaki, J., Kertész, J. and Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Phys. Rev. E* **71**, 065103.

Pal, C., Papp, B., Lercher, M. J., Csermely, P., Oliver, S. G. and Hurst, L. D. (2006). Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667-670.

Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S. and Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. *Trends. Biotechnol*. **22**, 400-405.

Papp, B., Pal, C. and Hurst, L. D. (2004). Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**, 661-664.

Pastor-Satorras, R., Vazquez, A. and Vespignani, A. (2001). Dynamical and correlation properties of the Internet. *Phys. Rev. Lett*. **87**, 258701.

Price, D. J. d. (1965). Networks of scientific papers. *Science* **149**, 510-515.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-1555.

Reed, J. L. and Palsson, B. O. (2004). Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome. Res*. **14**, 1797-1805.

Sauer, U., Lasko, D. R., Fiaux, J., Hochuli, M., Glaser, R., Szyperski, T., Wuthrich, K. and Bailey, J. E. (1999). Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. *J. Bacteriol*. **181**, 6679-6688.

Schilling, C. H., Letscher, D. and Palsson, B. O. (2000). Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol*. **203**, 229-248.

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S. and Palsson, B. O. (2002). Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol*. **184**, 4582-4593.

Schuster, S. and Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst*. **2**, 165-182.

Segre, D., DeLuna, A., Church, G. M. and Kishony, R. (2005). Modular epistasis in yeast metabolism. *Nat. Genet*. **37**, 77-83.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623-627.

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

Wagner, A. (2003). How the global structure of protein interaction networks evolves. *Proc. R. Soc. Lond. B Biol. Sci.* **270**, 457-466.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge: Cambridge University Press.

Watts, D. and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature* **393**, 440-442.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol*. **4**, 17.

The Journal of Experimental Biology 210, 1492-1496
Published by The Company of Biologists 2007
doi:10.1242/jeb.02783

Glossary of terms

This section is designed to help readers adapt to the complex terminology associated with contemporary molecular genetics, genomics and systems biology. Fuller descriptions of these terms are available at http://www.wikipedia.org/

Ab initio prediction	methods used to predict the potential genes encoded in the genome, which are trained on datasets made of known genes, and used computationally to predict coding regions out of genome without the aid of cDNA sequence. Although their performance is improving, these algorithms perform very poorly on non-protein coding genes.
Annotation	as applied to proteins, DNA sequences or genes. The storage of data describing these entities (protein/gene identities, DNA motifs, gene ontology categorisation, etc.) within a biological database. Active projects include FlyBase and WormBase. See Gene ontology.
Assembly	the process of aligning sequenced fragments of DNA into their correct positions within the chromosome or transcript.
cDNA	complementary DNA. This is DNA synthesised from a mature mRNA template by the enzyme reverse transcriptase. cDNA is frequently used as an early part of gene cloning procedures, since it is more robust and less subject to degradation than the mRNA itself.
ChIP	chromatin immunoprecipitation assay used to determine which segments of genomic DNA are bound to chromatin proteins, mainly including transcription factors.
Chip	see Microarray.
ChIP-on-chip	use of a DNA microarray to analyse the DNA generated from chromatin immunoprecipitation experiments (see ChIP).
cis-acting	a molecule is described as *cis*-acting when it affects other genes that are physically adjacent, on the same chromosome, or are genetically linked or in close proximity (for mRNA expression, typically a promoter).
Collision-induced dissociation	a mechanism by which molecules (e.g. proteins) are fragmented to form molecular ions in the gas phase. These fragments are then analysed within a mass spectrometer to provide mass determination.
Connectivity	a term from graph theory, which indicates the number of connections between nodes or vertices in a network. Greater connectedness between nodes is generally used as a measure of robustness of a network.
CpG islands	regions that show high density of ‘C followed by G’ dinucleotides and are generally associated with promoter elements; in particular, stretches of DNA of at least 200 bp with a C–G content of 50% and an observed CpG/expected CpG in excess of 0.6. The cytosine residues can be methylated, generally to repress transcription, while demethylated CpGs are a hallmark of transcription. CpG dinucleotides are under-represented outside regulatory regions, such as promoters, because methylated C mutates into T by deamination.
Edge	as in networks. Connects two nodes (or vertices) within a system. These concepts arise from graph theory.
Enhancer	a short segment of genomic DNA that may be located remotely and that, on binding particular proteins (*trans*-acting factors), increases the rate of transcription of a specific gene or gene cluster.
Epistasis	a phenomenon when the properties of one gene are modified by one or more genes at other loci. Otherwise known as a genetic interaction, but epistasis refers to the statistical properties of the phenomenon.

eQTL	the combination of conventional QTL analysis with gene expression profiling, typically using microarrays. eQTLs describe regulatory elements controlling the expression of genes involved in specific traits.
EST	expressed sequence tag. A short DNA sequence determined for a cloned cDNA representing portions of an expressed gene. The sequence is generally several hundred base pairs from one or both ends of the cloned insert.
Exaptation	a biological adaptation where the current function is not that which was originally evolved. Thus, the defining (derived) function might replace or persist with the earlier, evolved adaptation.
Exon	any region of DNA that is transcribed to the final (spliced) mRNA molecule. Exons interleave with segments of non-coding DNA (introns) that are removed (spliced out) during processing after transcription.
Gene forests	genomic regions for which RNA transcripts, produced from either DNA strand, have been identified without gaps (non-transcribed genomic regions). Conversely, regions in which no transcripts have ever been detected are called ‘gene deserts’.
Gene interaction network	a network of functional interactions between genes. Functional interactions can be inferred from many different data types, including protein–protein interactions, genetic interactions, co-expression relationships, the co-inheritance of genes across genomes and the arrangement of genes in bacterial genomes. The interactions can be represented using network diagrams, with lines connecting the interacting elements, and can be modelled using differential equations.
Gene ontology (GO)	an ontology is a controlled vocabulary of terms that have logical relationships with each other and that are amenable to computerised manipulation. The Gene Ontology project has devised terms in three domains: biological process, molecular function and cell compartment. Each gene or DNA sequence can be associated with these annotation terms from each domain, and this enables analysis of microarray data on groups of genes based on descriptive terms so provided. See http://www.geneontology.org
Gene set enrichment analysis	a computational method that determines whether a defined set of genes, usually based on their common involvement in a biological process, shows statistically significant differences in transcript expression between two biological states.
Gene silencing	the switching-off of a gene by an epigenetic mechanism at the transcriptional or post-transcriptional levels. Includes the mechanism of RNAi.
Genetic interaction (network)	a genetic interaction between two genes occurs when the phenotypic consequences of a mutation in one gene are modified by the mutational status at a second locus. Genetic interactions can be aggravating (enhancing) or alleviating (suppressing). To date, most high-throughput studies have focussed on systematically identifying synthetic lethal or sick (aggravating) interactions, which can then be visualised as a network of functional interactions (edges) between genes (nodes).
Genome	a portmanteau of gene and chromosome, the entire hereditary information for an organism that is embedded in the DNA (or, for some viruses, in RNA). Includes protein-coding and non-coding sequences.
Heritability	phenotypic variation within a population is attributable to the genetic variation between individuals and to environmental factors. Heritability is the proportion due to genetic variation usually expressed as a percentage.
Heterologous hybridization	the use of a cDNA or oligonucleotide microarray of probes designed for one species with target cRNA/cDNAs from a different species.
Homeotic	the transformation of one body part to another due to mutation of specific developmentally related genes, notably the *Hox* genes in animals and *MADS-box* genes in plants.
Hub	as in networks. A node with high connectivity, and thus which interacts with many other nodes in the network. A hub protein interacts with many other proteins in a cell.

Hybridisation	the process of joining (annealing) two complementary single-stranded DNAs into a single double-stranded molecule. In microarray analysis, the target RNA/DNA from the subject under investigation is denatured and hybridised to probes that are immobilised on a solid phase (i.e. glass microscope slide).
Hypomorph	in genetics, a loss-of-function mutation in a gene, but which shows only a partial reduction in the activity it influences rather than a complete loss (cf. hypermorph, antimorph, neomorph, etc).
Imprinting	a phenomenon where two inherited copies of a gene are regulated in opposite ways, one being expressed and the other being repressed.
Indel	insertion and deletion of DNA, referring to two types of genetic mutation. To be distinguished from a 'point mutation', which refers to the substitution of a single base.
Interactome	a more or less comprehensive set of interactions between elements within cells. Usually applied to genes or proteins as defined by transcriptomic, proteomic or protein–protein interaction data.
Intron	see Exon.
KEGG	The Kyoto Encyclopedia of Genes and Genomes is a database of metabolic and other pathways collected from a variety of organisms. See http://www.genome.jp/kegg
Metabolomics	the systematic qualitative and quantitative analysis of small chemical metabolite profiles. The metabolome represents the collection of metabolites within a biological sample.
Metagenomics	the application of genomic techniques to characterise complex communities of microbial organisms obtained directly from environmental samples. Typically, genomic tags are sequence characterised as markers of each species to inform on the range and abundance of species in the community.
Microarray	an arrayed set of probes for detecting molecularly specific analytes or targets. Typically, the probes are composed of DNA segments that are immobilised onto the solid surface, each of which can hybridise with a specific DNA present in the target preparation. DNA microarrays are used for profiling of gene transcripts.
Model species	a species used to study particular biological phenomena, the outcome offering insights into the workings of other species. Usually, the selection is based on experimental tractability, particularly ease of genetic manipulation. For the geneticist, it is an organism with inbred lines where sibs will be >98% identical (i.e. *Drosophila*, *Caenorhabditis elegans* and mice). For genomic science, it refers to a species for which the genomic DNA has been sequenced.
miRNA	a category of novel, very short, non-coding RNAs, generated by the cleavage of larger precursors (pri-miRNA). These short RNAs are included in the RNA-induced silencing complex (RISC) and pair to the 3′ ends of target RNA, blocking its translation into proteins (in animals) or promoting RNA cleavage and degradation (in plants).
mRNA	a protein-coding mRNA containing a protein-coding region (CDS), preceded by a 5′ and followed by a 3′ untranslated region (5′ UTR and 3′ UTR). The UTRs contain regulatory elements. A full-length cDNA contains the complete sequence of the original mRNA, including both UTRs. However, it is often difficult to assign the starting–termination positions for protein synthesis unambiguously. A cDNA containing the entire CDS is often considered acceptable for bioinformatic and experimental studies requiring full-length cDNAs.
ncRNA	non-coding RNA is any RNA molecule with no obvious protein-coding potential for at least 80 or 100 amino acids, as determined by scanning full-length cDNA sequences. It includes ribosomal (rRNA) and transfer RNAs (tRNA) and is now known to include various sub-classes of RNA, including snoRNA, siRNA and piRNA. Just like the coding mRNAs, a large proportion of ncRNAs are transcribed by RNA polymerase II and are large transcripts. A description of the many forms of ncRNA can be found at http://en.wikipedia.org/wiki/Non-coding_RNA.

Node	as in networks. Objects linked by edges to create a network.
PCR	polymerase chain reaction. A molecular biology technique for replicating DNA *in vitro*. The DNA is thus amplified, sometimes from very small amounts. PCR can be adapted to perform a wide variety of genetic manipulations.
piRNA	Piwi-interacting RNA. A class of RNA molecules (29–30 nt long) that complex with Piwi proteins (a class of the Argonaute family of proteins) and are involved in transcriptional gene silencing.
PMF	peptide mass fingerprinting. An analytical technique for protein identification in which a protein is fragmented using proteases. The resulting peptides are analysed by mass spectrometry and these masses compared against a database of predicted or measured masses to generate a protein identity.
Polyadenylation	the covalent addition of multiple A bases to the 3′ tail of an mRNA molecule. This occurs during the processing of transcripts to form the mature, spliced molecule and is important for regulation of turnover, trafficking and translation.
Post-source decay	in mass spectrometry. The fragmentation of precursor molecular ions as they accelerate away from the ionisation source of the mass spectrometer. All precursor ions leaving the ion source have approximately the same kinetic energy, but fragmentation results in smaller product ions that can be distinguished from precursor ions using a ‘reflectron’ by virtue of their lower kinetic energies.
Post-translational modification	the chemical modification of a protein after synthesis through translation. Some modifications, notably phosphorylation, affect the properties of the protein, offering a means of regulating function.
Principal component analysis (PCA)	a technique for simplifying complex, multi-dimensional datasets to a reduced number of dimensions, the principal components. This procedure retains those characteristics of the data that relate to its variance.
Promoter	a regulatory DNA sequence, generally lying upstream of an expressed gene, which in concert with other often distant regulatory elements directs the transcription of a given gene.
Proteome	the entire protein complement of an organism, tissue or cell culture at a given time.
Quantitative trait	inheritance of a phenotypic property or characteristic that varies continuously between extreme states and can be attributed to interactions between multiple genes and their environment.
qPCR	quantitative real-time PCR, sometimes called real-time PCR. A more quantitative form of RT-PCR in which the quantity of amplified product is estimated after each round of amplification.
QTL	quantitative trait loci. A region of DNA that contains those genes contributing to the trait under study.
RISC	RNA-induced silencing complex. A protein complex that mediates the double-stranded RNA-induced destruction of homologous mRNA.
RNAi	RNA interference or RNA-mediated interference. The process by which double-stranded RNA triggers the destruction of homologous mRNA in eukaryotic cells by the RISC.
RT-PCR	reverse transcription–polymerase chain reaction. A technique for amplifying a defined piece of RNA that has been converted to its complementary DNA form by the enzyme reverse transcriptase. See qPCR.
siRNA	small interfering RNA, or silencing RNA. A class of short (20–25 nt), double-stranded RNA molecules. It is involved in the RNA interference pathway, which alters RNA stability and thus affects RNA concentration and thereby suppresses the normal expression of specific genes. Widely used in biomedical research to ablate specific genes.

snoRNA	small nucleolar RNA. A sub-class of RNA molecules involved in guiding chemical modification of ribosomal RNA and other RNA genes as part of the regulation of gene expression.
SNP	single nucleotide polymorphism. A single base-pair mutation at a specific locus, usually consisting of two alleles. Because SNPs are conserved over evolution, they are frequently used in QTL analysis and in association studies in place of microsatellites, and in genetic fingerprinting analyses.
SSH	suppressive subtractive hybridisation. A powerful protocol for enriching cDNA libraries for genes that differ in representation between two or more conditions. It combines normalisation and subtraction in a single procedure and allows the detection of low-abundance, differentially expressed transcripts, such as those involved in signalling and signal transduction.
Structural RNAs	a class of non-coding RNA, long known to have a structural role (for instance, the ribosomal RNAs), transcribed by RNA polymerase I or III.
Systems biology	treatment of biological entities as systems composed of defined elements interacting in defined ways to enable the observed function and behaviour of that system. The properties of the systems are embedded in a quantitative model that guides further tests of systems behaviour.
TATA-boxes	sequences in promoter regions constituted by TATAAA, or similar variants, which were considered the hallmark of Promoters. Recent data show that they are present only in the minority of promoters, where they direct transcription at a single well-defined location some 30 bp downstream of this element.
trans-acting	a factor or gene that acts on another unlinked gene, a gene on a separate chromosome or genetically unlinked usually through some diffusible protein product (for mRNA expression, typically a transcription factor).
Transcript	an RNA product produced by the action of RNA polymerase reading the sequence of bases in the genomic DNA. Originally limited to protein-coding sequences with flanking UTRs but now known to include large numbers of products that do not code for a protein product.
Transcriptome	the full set of mRNA molecules (transcripts) produced by the system under observation. Whilst the genome is fixed for a given organism, the transcriptome varies with context (i.e. tissue source, ontogeny, external conditions or experimental treatment).
Transgene	a gene or genetic material that has been transferred between species or between organisms using one of several genetic engineering techniques.
Transinduction	generation of transcripts from intergenic regions. At least some such products do not relate to a definable promoter or transcriptional start site.
Transposon	sequences of DNA able to move to new positions within the genome of a single cell. This event might cause mutation at the site of insertion. Also called ‘mobile genetic elements’ or ‘jumping genes’.
Transvection	an epigenetic phenomenon arising from the interaction between one allele and the corresponding allele on the homologous chromosome, leading to gene regulation.
TUs	transcriptional units. Used to group all of the overlapping RNA transcripts that are transcribed from the same genomic strand and share exonic sequences.
UTR	untranslated region. Regions of the mRNA that lie at either the 3′ or 5′ flanking ends of the molecule (i.e. 3′ UTR and 5′ UTR). They bracket the protein-coding region and contain signals and binding sites that are important for the regulation of both protein translation and RNA degradation.