The Architecture of Biological Networks

Stefan Wuchty[†], Erszébet Ravasz[†] and Albert-László Barabási[†,*]

[†]Department of Physics

225 Nieuwland Science Hall

University of Notre Dame

Notre Dame, IN 46556

USA

[*]Address for correspondence.

Tel: ++1 574 631 5767, Fax: ++1 574 631 5259, Email: `alb@nd.edu`

INTRODUCTION

Understanding complex systems often requires a bottom-up approach, breaking the system into small and elementary constituents and mapping out the interactions between these components. In many cases, the myriads of components and interactions are best characterized as networks. For example, the society is a network of people connected by various links, including friendships (Milgram, 1967), collaborationships (Kochen, 1989; Wasserman & Faust, 1994), sexual contacts (Liljeros et al., 2001) or scientific co-authorships (Redner, 1998; Newman, 2001). Electronic communication relies on two very different networks: the physical network wiring the routers together (Internet) (Faloutsos, Faloutsos & Faloutsos, 1999; Vázquez, Pastor-Satorras & Vespignani, 2002) and the web of homepages linked by URLs (World Wide Web) (Albert, Jeong & Barabási, 1999; Lawrence & Giles, 1999; Broder et al., 2000). Airline, cell-phone, power-grid or business networks represent further examples of complex networks of technological, scientific or economic interest.

In biological systems networks emerge in many disguises, from food webs in ecology to various biochemical nets in molecular biology. In particular, the wide range of interactions between genes, proteins and metabolites in a cell are best represented by various complex networks. During the last decade, genomics has produced an incredible quantity of molecular interaction data, contributing to maps of specific cellular networks. The emerging fields of transcriptomics and proteomics have the potential to join the already extensive data sources provided by the genome wide analysis of gene expression at the mRNA and protein level (Pandey & Mann, 2000; Caron et al., 2001; Burge, 2001). Indeed, extensive protein-protein interaction maps have been generated for a variety of organisms including viruses (Flajolet et al., 2000; McGraith et al., 2000), prokaryotes, like *H.pylori* (Rain et al., 2001) and eukaryotes, like *S. cerevisiae* (Ito et al., 2000; Ito et al., 2001; Schwikowski, Uetz & Fields, 2000; Uetz et al., 2000; Gavin et al., 2002; Ho et al., 2002; Jeong et al., 2001)

and *C.elegans* (Walhout et al., 2000). Beyond the current focus on uncovering the structure of genomes, proteomes and interactomes of various organisms, some of the most extensive datasets are the metabolic maps (Overbeek et al., 2000; Karp et al., 2000), catalyzing an increasing number of studies focusing on the architecture of the metabolism (Jeong et al., 2000; Fell & Wagner, 2000; Wagner & Fell, 2001).

Networks offer us a new way to categorize systems of very different origin under a single framework. This approach has uncovered unexpected similarities between the organization of various complex systems, indicating that the networks describing them are governed by generic organization principles and mechanisms. Understanding the driving forces which invest different networks with similar topological features enables systems biology to combine the numerous details about molecular interactions into a single framework, offering means to address the structure of the cell as a whole.

BASIC NETWORK FEATURES

A node's degree (or connectivity), giving the number of links k the node has, is the most elementary network measure. For example, in Fig. 1 nodes i and j have exactly three links ($k = 3$). The overall graph, however, is characterized by the average degree, $\langle k \rangle$, which has the value $\langle k \rangle = 2.6$ for this example. Yet, the average degree does not capture the potential degree variations present in the network. This is better characterized by the degree distribution, $P(k)$, which gives the number of nodes with exactly k links.

Planing a trip from Anchorage, Alaska to Alice Springs in the outbacks of Australia requires finding the shortest paths through a particular airline's transportation network. As in most networks, there are multiple paths between any two nodes i and j, a useful distance measure is the length of the shortest path, l_{ij} (see Fig. 1). The

mean path length defined as

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

offering a measure of the network's navigability. A network which can be 'crossed' by a relatively small number of steps is often referred to display the 'small world' property, first illustrated on social networks, indicating that two randomly chosen individuals can be connected by only six intermediate acquaintances (Milgram, 1967).

Nodes in many real systems exhibit a tendency to cluster, which can be quantified using the clustering coefficient (Watts & Strogatz, 1998), a measure of the degree to which the neighbors of a particular node are connected to each other (Fig. 2). For example, in a friendship network C reflects the degree to which friends of a particular person are friends with each other as well. Formally, the clustering coefficient of node i is defined as

$$C_i = \frac{2n_i}{k_i(k_i-1)}, \tag{2}$$

where n_i denotes the number of links connecting the k_i neighbors of node i to each other. Accordingly, we can define the average clustering coefficient as

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{3}$$

An additional important measure of the network's structure is the function $C(k)$, defined as the average clustering coefficient of all nodes with k links. If $C(k)$ is independent of k the network is either homogeneous or it is dominated by numerous small tightly linked clusters. In contrast, if $C(k)$ follows $C(k) \sim k^{-1}$, the network has a hierarchical architecture meaning that sparsely connected nodes are part of highly clustered areas (Ravasz et al., 2002; Ravasz & Barabási, 2002; Dorogovtsev, Goltsev & Mendes, 2002; Jung, Kim & Kahng, 2002). In such hierarchical networks, communication between the different highly clustered neighborhoods are maintained by a few hubs.

As we will see below, the degree distribution $P(k)$ and the k dependence of $C(k)$ can have generic features, allowing us to classify various network. Parameters such as the average degree $\langle k \rangle$, average path length $\langle l \rangle$ and average clustering coefficient $\langle C \rangle$ characterize the unique properties of the particular network under consideration, and therefore are less generic.

NETWORKS MODELS

The main role of the network models is to explain the emergence and behavior of some of the most important network characteristics. As they play a crucial role in shaping our understanding of complex networks, we need to pay attention to some of the more important models.

Random Networks

While graph theory initially focused on regular graphs, since the 1950's large networks with no apparent design principles were described as random graphs (Bollobás, 1985), proposed as the simplest and most straightforward realization of a complex network. According to the Erdős-Rényi (ER) model of random graphs (Erdös & Rényi, 1960), we start with N nodes and connect each pair of nodes with probability p, creating a graph with approximately $pN(N-1)/2$ randomly distributed links (first column in Fig. 3). The ER graph has an exponential degree distribution and exhibits the small-world property. Indeed, in the ER network, most nodes have approximately the same number of links, $k \approx \langle k \rangle$ (first column in Fig. 4), and the mean path length is proportional to the network size, $L \sim \log N$.

The growing interest in complex systems prompted many scientists to ask a simple question: are real networks behind diverse complex systems, like the cell, fundamentally random?

Scale-free Networks

A highly nontrivial development in our understanding of complex networks was the discovery that for most large networks, including the metabolic and protein interaction networks (Jeong et al., 2000; Jeong et al., 2001), the degree distribution follows a power-law

$$P(k) \sim k^{-\gamma}. \tag{4}$$

These networks are called scale-free, as a power-law does not support the existence of a characteristic scale. Two mechanisms, absent from the classical random network model, are responsible for the emergence of this power-law degree distribution (Barabási & Albert, 1999; Barabási, Albert & Jeong, 1999). First, most networks grow through the addition of new nodes, that link to nodes already present in the system. Second, in most real networks there is a higher probability to link to a node with a large number of connections, a property called preferential attachment. The scale-free model introduced by Barabási and Albert (BA) (second columns in Fig. 3 and 4) incorporates these features. Starting from a small graph, at each time step a node with m links is added to the network, connecting to a previously present node i with probability

$$\Pi_i = k_i / \sum_j k_j \,, \tag{5}$$

where k_i is the degree of node i. The network generated by this growth process will be scale-free with degree exponent $\gamma = 3$. In a scale-free network the probability that a node is highly connected ($k \gg \langle k \rangle$) is statistically more significant than in a random graph. Thus, the network's properties are often determined by a relatively small number of highly connected nodes or hubs. An important consequence of the hubs is that scale-free networks exhibit high tolerance to random perturbations but are sensitive to targeted attack on the highly connected nodes (Albert, Jeong & Barabási, 2000). Accordingly, failure of randomly selected nodes cannot destroy the network's integrity. However, the systematic removal of the hubs will rapidly fragment the network. This feature is of particular importance for biological systems, since it

reflects the biochemical network's resilience against random mutations. Therefore, highly connected nodes in biochemical networks might be potential candidates for drug targets.

The presence of hubs in a scale-free network has a fundamental impact on virus spreading as well. Classical epidemiological models predict that infectious diseases with transmission probability under an epidemic threshold will inevitably die out. However, in scale-free networks the epidemic threshold is reduced to zero (Pastor-Satorras & Vespignani, 2001). Thus, as some social and sexual networks are known to exhibit a scale-free topology (Liljeros et al., 2001), even extremely weakly infectious viruses can spread and prevail, making random immunization ineffective.

Hierarchical Networks

Many real networks are expected to be fundamentally modular, meaning that the network can be seamlessly partitioned into a collection of modules. Each module is expected to perform an identifiable task, separable from the function of other modules (Hartwell et al., 1999; Wolf, Karev & Koonin, 2002; Lauffenburger, 2000; Shen-Orr et al., 2002). Therefore, we must reconcile the scale-free property with the network's potential modularity. Numerical simulations indicate that neither the random nor the scale-free network model are modular.

In order to account for the coexistence of modularity, local clustering and scale-free topology in real systems, we have to assume that clusters combine in an iterative manner, generating a hierarchical network (Ravasz & Barabási, 2002; Barabási, Ravasz & Vicsek, 2001). Such networks emerge from an iterative duplication and integration of clustered nodes, a process which in principle can be repeated indefinitely. Our starting point is a small cluster of four densely linked nodes. Next we generate three replicas of this hypothetical module and connect the three external nodes of the replicated clusters to the central node of the old cluster, obtaining a large 16-node

module. Subsequently, we again generate three replicas of this 16-node module, and connect the 16 peripheral nodes to the central node of the old module, obtaining a new module of 64 nodes (third column of Fig. 3).

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 $\gamma = 1 + \ln 4/\ln 3 = 2.26$. Yet, the most important signature of this hierarchical modularity is the fact that the clustering coefficient, $C(k)$, scales as k^{-1} (third column of Fig. 4). Note, that for the network generated by the ER and BA models $C(k)$ is independent of k.

Modularity does not, however, imply clear-cut subnetworks which are linked in well-defined ways. In fact, the boundaries of modules are often considerably blurred, triggered by highly connected nodes which interconnect modules.

BIOLOGICAL NETWORKS

Metabolic Networks

The structure of metabolic networks was addressed by two independent studies by Fell and Wagner and Jeong *et al.* Fell and Wagner assembled a list of stoichiometric equations that represent the central routes of the energy metabolism and small-molecule building block synthesis in *E.coli* (Fell & Wagner, 2000; Wagner & Fell, 2001). A substrate graph was defined by the nodes representing all metabolites, two substrates being considered linked if they occurred in the same reaction. They found the substrate graph to be scale-free with glutamate, coenzyme A, 2-oxoglutarate, pyruvate and glutamine having the highest degree which were viewed as an evolutionary core of the *E.coli*.

At the same time, Jeong *et al.* analyzed the metabolic networks of 43 organisms representing all three domains of life (Jeong et al., 2000), finding that the power-law degree distribution for both incoming and outgoing edges holds for organisms of all

kingdoms. Furthermore, the average separation between nodes has the same value for all organisms under consideration, regardless of the number of substrates found in the given species. Interestingly, the ranking of the most connected substrates is largely identical for all organisms. A recent study comparing the system-level properties of metabolic networks in various organisms indicates that the structural features of these networks are more conserved than the components themselves (Podani et al., 2001; Wolf, Karev & Koonin, 2002).

Protein Interaction Networks

Protein interactions offer another opportunity to study cellular networks, considering proteins as nodes and physical interactions (binding) as links. It has been shown that interaction networks of *S. cerevisiae* and *H.pylori* proteins exhibit distinct scale-free behavior (Jeong et al., 2001; Wagner, 2001). Although protein interaction data is derived from different sources and is retrieved by different methods, the emergence of the scale-free property appears to be a robust feature. As previously discussed, scale-free networks are vulnerable upon targeted attack on their highly connected nodes. Therefore, mutations of highly interacting proteins are expected to be lethal for the cell. This prediction is supported by explicit measurements (Jeong, Oltvai & Barabási, 2003). Fig. 5 represents the Yeast protein interaction network, illustrating the basic feature that hubs keep many sparsely nodes together.

Protein Domain Networks

The domain architecture of proteins was studied by considering protein domains as nodes and their co-occurrence in proteins as links (Wuchty, 2001; Apic, Gough & Teichmann, 2001; Wuchty, 2002), documenting again the emergence of a scale-free architecture. Although methods and sources of domain information were different, the scale-free features of the networks were found to be robust. Domains which appear in cellular functions crucial for the maintenance of multi-cellular organisms, such

as signal transduction and cell-cell contacts, were found to be the most connected. Thus, domains like kinases, immunoglobulins and zinc-fingers played an important role. Interestingly, the increasing complexity of organism's domain architecture was found to decrease the slope of the degree distribution and highly connected domains constantly accumulated links due to the organismic complexity. Similarly, interactions of domain families generated from sequence and structural data (Park, Lappe & Teichmann, 2001; Wuchty, 2002) revealed that highly connected domains on sequence level appear to be the most frequently interacting as well.

Hierarchies in Biological Networks

The clustering coefficient of metabolic networks varies with the inverse degree, $C(k) \sim k^{-1}$, indicating the presence of a hierarchical modularity. In order to discern the discrete modules, we can define a topological overlap, which scales from 0 to 1, reflecting the degree to which two metabolites i and j interact with the same substrates. Substrates that are part of larger metabolic modules appear to have a high topological overlap with their neighbors. The application of average-linkage clustering to the obtained overlap matrix has been used to uncover the topological modules present in the metabolism (Fig. 6). The clustering identified a hierarchy of nested topological modules of increasing sizes and decreasing interconnectedness. The hierarchical tree offers a breakdown of the metabolism into several large modules which are further partitioned into smaller but more integrated submodules, reflecting a certain degree of inherent self-similarity. Some of these modules have been found to be in excellent agreement with the known functional classification of metabolites. Other approaches to discern modules in metabolic networks focused on the appearance of edges in mutual shortest paths in the network (Holme, Huss & Jeong, 2003; Girvan & Newman, 2002). The most frequent edges were identified and removed in an iterative manner, uncovering again the underlying functional modules.

Finally, modularity is not an exclusive property of the metabolism. Indeed, the protein interaction network of *S. cerevisae* (Yook, Oltvai & Barabási, 2003), based on four independent databases (Xenarios et al., 2001; Mewes et al., 2000; Uetz et al., 2000; Ito et al., 2001) and the conformational spaces of RNA (Wuchty, 2003) also reflect a modular architecture.

Mechanisms of Proteome Evolution

The origin of the scale-free behavior in biological networks continues to offer some unresolved questions. Recently, however, it has been shown that a simple model based on gene duplication leads to the experimentally observed scale-free topology of protein-protein interaction networks (Wagner, 2001; Vazquez et al., 2003; Solé et al., 2002; Pastor-Satorras, Smith & Solé, 2002). In the model, at each time step a gene is randomly chosen and duplicated. The copied gene retains all interactions of the original gene. To mimic the potential loss or gain of interactions due to random mutations, interactions of the duplicated genes are deleted or newly added with probabilities δ and α, respectively (Fig. 7). The emerging network can be shown analytically to have a power-law degree distribution, high clustering coefficient and a visual structure similar to the protein-protein interaction network shown in Fig. 5.

CONCLUSIONS

The power-law degree distribution, the quantitative signature of a scale-free network, has emerged as one of the few universal laws characterizing cellular networks. Of even greater immediate importance is the intriguing possibility of using the insights provided by the scale-free models as a framework to facilitate the analysis of biological networks at a higher level of abstraction. Such approaches could reveal salient features of biological phenomena missed by non-network based approaches.

The appearance of hierarchical modularity in biological networks supports the assumption that evolution acts on many levels. The accumulation of local changes, affecting the small highly integrated modules, slowly impacts the larger, less integrated modules as well. Thus, evolution might act in self-similar fashion, copying and reusing existing modules to further increase the organism's complexity. Especially in the face of eukaryotic evolution, this network based framework might be suitable to describe the explosion of complexity in the development of the single-celled *S. cerevisiae* toward the multicellular *H. sapiens*.

It is widely accepted that different cellular functions, such as information storage, processing and execution is carried out by the genome, transcriptome, proteome and metabolome. Although the functional distinction between these organizational levels is not always clear cut since e.g. the proteome is crucial for short term information storage, all cellular functions can be described by networks of various heterogeneous components. One way to visualize the complex relationships between these components is to organize them into a simple complexity pyramid (Oltvai & Barabási, 2002) in which various molecular components - genes, RNAs, proteins and metabolites - organize themselves into recurrent patterns such as metabolic pathways and genetic regulatory motifs. In turn, motifs and pathways are seamlessly integrated to form functional modules which are responsible for distinct cellular functions (Hartwell et al., 1999). These modules are nested in a hierarchical fashion and define the cell's large-scale organization (Fig. 8).

Our present knowledge about the architecture of biological networks emphasizes two major aspects: (1) Discrete cellular functions are mediated with the aid of distinct albeit often blurred modules; (2) Network integrity is assured by a handful highly connected nodes, making networks robust against random failures but exceedingly vulnerable upon targeted attack. These features explain the observation that many mutations have little or no phenotypic effect (Wagner, 2000) which appears to be

consistent with the presence of genes that either cannot propagate their failure or whose function can be replaced by other components of the network. The presence of genes that integrate multiple signals and can trigger widespread changes upon their failure proves the crucial role of highly connected genes.

For example, the tumor suppressor gene p53 has been identified as such a highly connected and thus crucial node which, once mutated, severely jeopardizes genome stability and integration of signals related to the control of cell-cycle and apoptosis (Vogelstein, Lane & Levine, 2000; Kohn, 1999). Emphasizing its crucial role, disfunctional p53 proteins are involved in more than half of all human cancer phenotypes. From a biomedical point of view, highly connected proteins in general and proteins which maintain the integrity of modules can be perceived as disease factors and thus potential drug targets. With the increasing ability to identify and collect protein-protein interactions the determination of modules and highly connected proteins will become a major issue in the fast and effective identification of potential drug targets.

The recent progress in biological networks has successively uncovered the skeleton and organization of networks, offering important insights about the assembly and functionality of components and subnetworks. In future, we will need to go several steps further addressing the dynamic aspects of various cellular networks. Especially, the analysis of fluxes and fluctuations along the links in metabolic and regulatory pathways will play a major role, significantly influencing potential biotechnological applications.

References

A. Vázquez, R. Pastor-Satorras & A. Vespignani (2002). Large-scale topological and dynamical properties of the Internet. *Phys. Rev. E, 65*, 066130.

Albert, R., Jeong, H. & Barabási, A.-L. (1999). Diameter of the World-Wide Web. *Nature, 401*, 130–131.

Albert, R., Jeong, H. & Barabási, A.-L. (2000). Attack and error tolerance of complex networks. *Nature, 406*, 378.

Apic, G., Gough, J. & Teichmann, S. (2001). Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes. *J. Mol. Biol., 310*, 311–325.

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

Barabási, A.-L., Albert, R. & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A, 272*, 173–187.

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

Bollobás, B. (1985). *Random Graphs.* Academic Press, London.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajalopagan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the web. *Comput. Netw., 33*, 309–320.

Burge, C. (2001). Chipping away at the transcriptome. *Nature Genet., 27*, 232–234.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.-C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S. & coauthors (2001). The Human Transcriptome Map: Clustering of Highly Expressed Genes in Chromosomal Domains. *Science, 291*, 1289–1292.

Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. (2002). Pseudofractal Scale-free Web. *Phys. Rev. E, 65*, 066122.

Erdös, P. & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci., 5*, 17–61.

Faloutsos, M., Faloutsos, P. & Faloutsos, C. (1999). On power-law relationships of the Internet topology. *Comput. Commun. Rev., 29*, 251–262.

Fell, D. & Wagner, A. (2000). The small world of metabolism. *Nature Biotech., 189*, 1121–1122.

Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., Tiollais, P., Transy, C. & Legrain, P. (2000). A genomic approach to the hepatitis c virus. *Gene, 242*, 369–379.

Gavin, A., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A.-M., Cruciat & coauthors (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature, 415*, 141–147.

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

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

Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutillier, K. & coauthors (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature, 415*, 180 – 183.

Holme, P., Huss, M. & Jeong, H. (2003). Subnetwork hierarchies in biochemical pathways. *Bioinformatics, 19*, 532–538.

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

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. & Sakaki, Y. (2000). Towards a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Nat. Acad. Sci. USA, 97*, 1143–1147.

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

Jeong, H., Oltvai, Z.N. & Barabási, A.-L. (2003). Prediction of protein essentiality based on genomic data. *ComPlexUs, 1*, 19–28.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature, 407*, 651–654.

Jung, S., Kim, S. & Kahng, B. (2002). A Geometric Fractal Growth Model for Scale Free Networks. *Phys. Rev. E, 65*, 056101.

Karp, P. D., Riley, M., Saier, M., Paulsen, I., Paley, S. & Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucl. Acids Res., 28*, 56–59.

Kochen, M., ed. (1989). *The Small World*. Ablex, Norwood, NJ.

Kohn, K. (1999). Molecular interaction map of mammalian cell-cycle control and DNA repair systems. *Mol. Biol. Cell, 10*, 2703–2734.

Lauffenburger, D. (2000). Cell signaling pathways as control modules: Complexity for simplicity. *Proc. Natl. Acad. Sci. USA, 97*, 5031–5033.

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the web. *Nature, 400*, 107–109.

Liljeros, F., Edling, C., Amaral, L. & Aberg, Y. (2001). The web of human sexual contacts. *Nature, 411*, 907–908.

McGraith, S., Holtzman, T., Moss, B. & Fields, S. (2000). Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc. Natl. Acad. Sci. USA, 97*, 4879–4884.

Mewes, H., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schüller & coauthors (2000). MIPS: a database for genomes and protein sequences. *Nucl. Acids Res., 28*, 37–40.

Milgram, S. (1967). The Small-World Problem. *Psychology Today, 2*, 60–67.

Newman, M. (2001). The structure of scientific collaboration networks. *Proc. Nat. Acad. Sci., 98*, 404–409.

Oltvai, Z.N. & Barabási, A.-L. (2002). Life's Complexity Pyramid. *Science, 298*, 763–764.

Overbeek, R., Larsen, N., Pusch, G., D'Souza, M., Selkov Jr, E., Kyrpides, N., Fonstein, M., Maltsev, N. & Selkov, E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res., 28*, 123–125.

Pandey, A. & Mann, M. (2000). Protemics to study genes and genomes. *Nature, 405*, 837 – 846.

Park, J., Lappe, M. & Teichmann, A. (2001). Mapping Protein Family Interations: Intramolecular and Intermolecular Protein Family Interaction Repertoires in the PDB and Yeast. *J. Mol. Biol., 307*, 929–938.

Pastor-Satorras, R., Smith, E. & Solé, R. (2002).. Evolving protein interaction networks through gene duplication. Santa Fe Working paper, 02-02-008.

Pastor-Satorras, R. & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Phys. Rev. Lett., 86*, 3200–3203.

Podani, J., Oltvai, Z.N., Jeong, H., Tombor, B., Barabaśi, A.-L. & Szathmary, E. (2001). Comparable system-level organization of Archae and Eukaryotes. *Nature Genet., 29*, 54–56.

Rain, J.-C., Selig, L., DeReuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter & coauthors (2001). The protein-protein interaction map of Helicobacter pylori. *Nature, 409*, 211–215.

Ravasz, E. & Barabási, A.-L. (2002). Hierarchical Organization in Complex Networks. *Phys. Rev. E, 67*, 026122.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science, 297*, 1551–1555.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Phys. J., B4*, 131135.

Schwikowski, B., Uetz, P. & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechn., 18*, 1257–1261.

Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. (2002). Network motifs in the transcriptional regulation network of E.coli. *Nature Genet., 31*, 64 – 68.

Solé, R., Pastor-Satorras, R., Smith, E. & Kepler, T. (2002). A model of large-scale proteome evolution. *Adv. Compl. Sys., 5*, 43–54.

Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshorn, D., Narayan, V., Srinivasan, M., Pochart, P. & coauthors (2000). A comprehensive analysis of protein-protein interactions of saccharomyces cerevisiae. *Nature, 403*, 623–627.

Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003). Modeling of protein interaction networks. *ComPlexUs, 1*, 38–44.

Vogelstein, B., Lane, D. & Levine, A. (2000). Surfing the p53 network. *Nature, 408*, 307–310.

Wagner, A. (2000). Mutational robustness in genetic networks of yeast. *Nat. Genet., 24*, 355 – 361.

Wagner, A. (2001). The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes. *Mol. Biol. Evol, 18*, 1283–1292.

Wagner, A. & Fell, D. A. (2001). The small world inside large metabolic networks. *Proc. Roy. Soc. London Series B, 268*, 1803 – 1810.

Walhout, A., Sordella, R., Lu, X., Hartley, J., Temple, G., Brasch, M., Thierry-Mieg, N. & Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science, 287*, 116–122.

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

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

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

Wuchty, S. (2001). Scale-Free Behavior in Protein Domain Networks. *Mol. Biol. Evol., 18*, 1694 – 1702.

Wuchty, S. (2002). Interaction and Domain Networks of Yeast. *Proteomics, 2*, 1715–1723.

Wuchty, S. (2003). Small-worlds in RNA. *Nucl. Acids Res., 31*, 1108–1117.

Xenarios, I., Fernandez, E., Salwinski, L., Duan, X., Thompson, M., Marcotte, E. & Eisenberg, D. (2001). DIP: the Database of Interacting Proteins: 2001 update. *Nucl. Acids Res., 29*, 239–241.

Yook, S. H., Oltvai, Z.N. & Barabási, A.-L. (2003). Functional and topological characterization of protein-protein interaction networks. submitted.

Fig. 1: Characterising a simple network: in the figure, both nodes, i and j, have three links ($k = 3$). The shortest path between these nodes, indicated in blue, has length $l_{ij} = 3$.

Fig. 2: The clustering coefficient C offers a measure of the degree of interconnectivity in the neighborhood of a node. For example, a node whose neighbors are all connected to each other has $C = 1$ (left), whereas a node with no links between its neighbors has $C = 0$ (right).

Fig. 3: **(a)** The random network model is constructed by laying down N nodes and connecting each pair of nodes with probability p. The figure shows a particular realization of such a network for $N = 10$ and $p = 0.2$. **(b)** The scale-free model assumes that the network constantly grows by the addition of new nodes. The figure shows the network at time t (nodes connected by green links) and after the addition of a new node at time $t + 1$ (red links). With the introduction of new nodes, already highly connected ones are more favored to be connected to the new one than less connected nodes. This procedure is called *preferential attachment*. **(c)** The iterative construction of a hierarchical network starts from a fully connected cluster of four nodes (blue), which is replicated three times. Subsequently, the peripheral nodes of each replica (green) are connected to the central node of the original module. Repeating the replication and the connection step with the 16-node module (red) leads to 64-nodes network which provides scale-free topology and is built by nested modules. **(d)** The random network is rather homogeneous, i.e. most nodes have approximately the same number of links. **(e)** In contrast, a scale-free network is extremely inhomogeneous: while the majority of nodes has one or two links, a few nodes have a large number of links preserving the systems integrity. To show this, five nodes with the highest number of links are colored red, and their first neighbors are colored green.

While in the random network only 27% of the nodes are reached by the five most connected nodes, in the scale-free network more than 60% are, demonstrating the key role hubs play in the scale-free network. Note, that both networks contain the same number of nodes and links. **(f)** A hierarchical network still preserves it's scale-free organization and displays inherent modularity of nodes. The node's affiliation to a certain module is indicated by different colors. However, the underlying network's structure clearly indicates blurred boundaries of it's modules.

Fig. 4: **(a)** For the random graph, the degree distribution, $P(k)$, which gives the probability that a randomly selected node has exactly k edges, follows a Poisson distribution which is strongly peaked at the average degree $\langle k \rangle$ and decays exponentially for large k. **(b), (c)** $P(k)$'s of a scale-free and a hierarchical network do not have a peak and decay as a power-law, $P(k) \sim k^{-\gamma}$. **(d), (e)** For both the random and the scale-free network, the $C(k)$ function, which denotes the mean clustering coefficient for nodes with exactly k links, is independent of k. **(f)** In contrast, $C(k)$ of a hierarchical network depends on k, decaying as $C(k) \sim k^{-1}$. Insets correspond to the number of the underlying networks.

Fig. 5: Map of the protein-protein interaction network of *S. cerevisiae* (Jeong et al., 2001). The color code of nodes refers to the phenotypic effect the deletion of the respective protein has on the organism (red: lethal, green: viable, orange: slowed growth, yellow:unknown).

Fig. 6: Hierarchies of topological modules in the *E.coli* metabolism. The branches of the tree obtained by average-linkage clustering of the topological overlap of metabolites (Ravasz et al., 2002) are color-coded to reflect the predominant biochemical

classification of their substrates. The biochemical classes represent carbohydrate metabolism (blue); nucleotide and nucleic acids metabolism (red); protein, peptide and amino acid metabolism (green); lipid metabolism (cyan); aromatic compound metabolism (dark pink); monocarbon metabolism (yellow) and coenzym metabolism(light orange) (Overbeek et al., 2000).

Fig. 7: Mechanism of the gene duplication and divergence model: At each time step, a gene is randomly duplicated retaining all of it's links (blue nodes and edges). Subsequently, interactions of the duplicated gene are deleted or newly added with probabilities δ and α, respectively (green edge).

Fig. 8: From the *particular* to the *universal*: The bottom (level 1) of the pyramid shows schematic representation of of the cell's functional organization: genome, transcriptome, proteome and metabolom. Insights into the cell's organization can be obtained if we consider the components to be linked by functional relationships, such as regulatory motifs and metabolic pathways (level 2). In turn, they are the building blocks of operational modules (level 3) which are nested and considerably blurred, generating a scale-free hierarchical architecture (level 4). Although the individual components are unique, the topological properties of biological networks share astounding similarities. This suggests that universal organizing principles apply to all kinds of complex networks (Oltvai & Barabási, 2002).

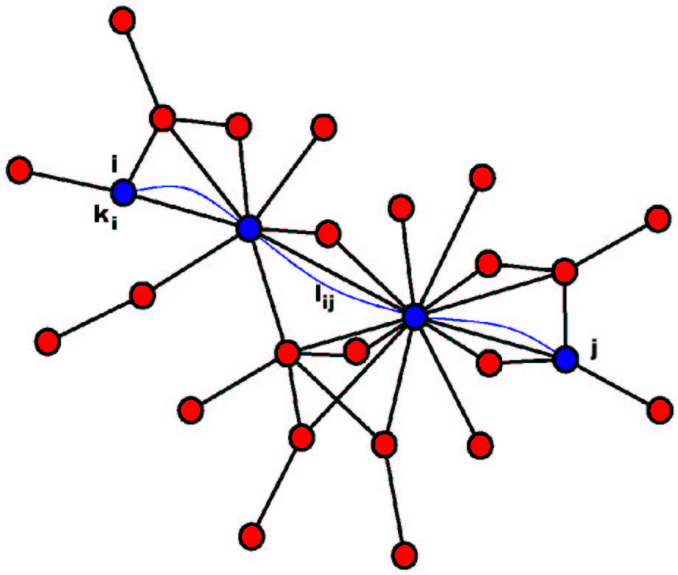

Figure 1

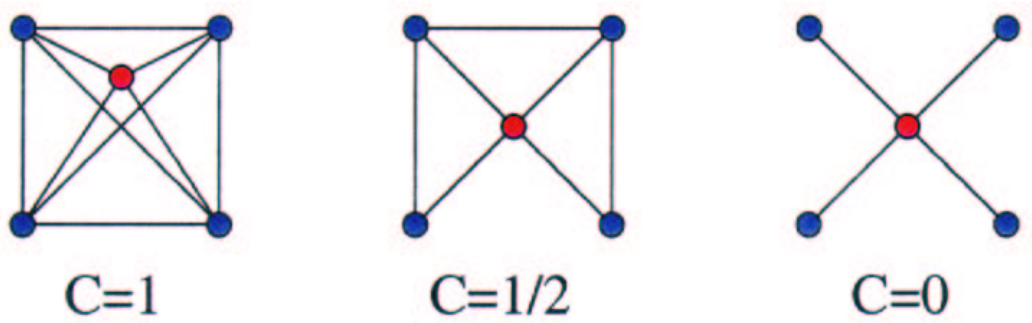

Figure 2

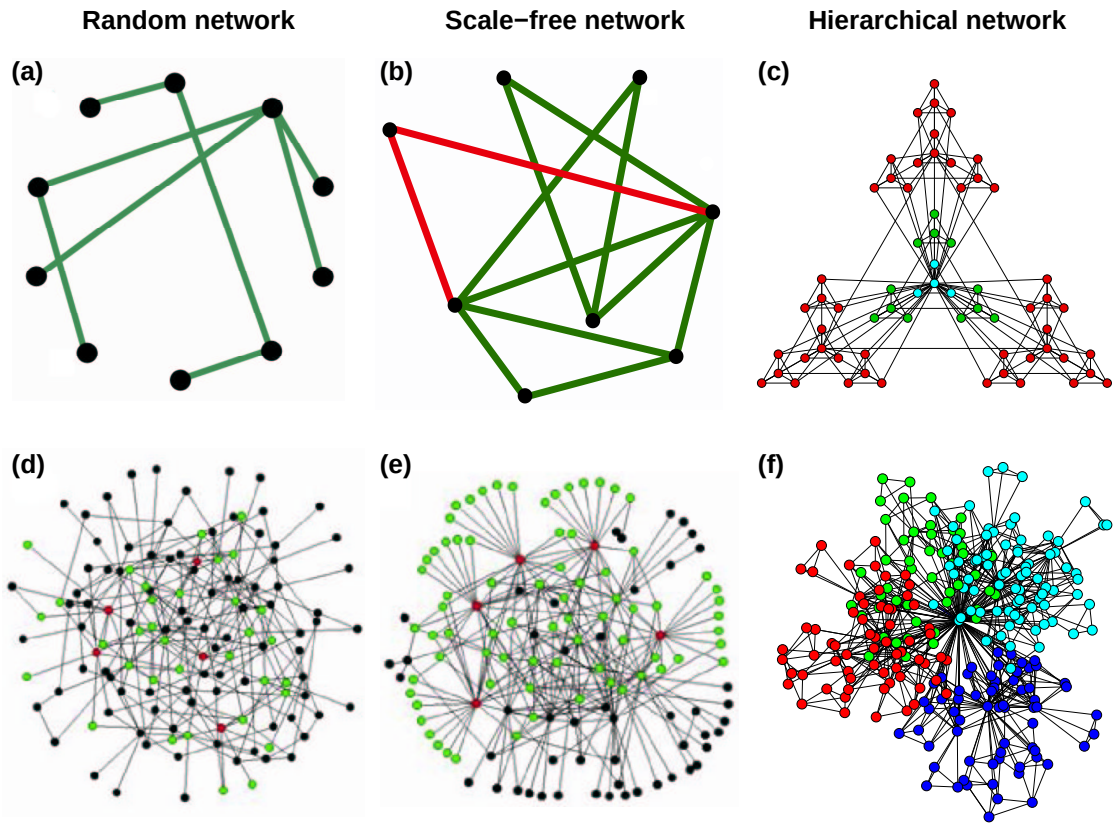
Random network
Scale−free network
Hierarchical network
(a)
(b)
(c)
(d)
(e)
(f)

Figure 3

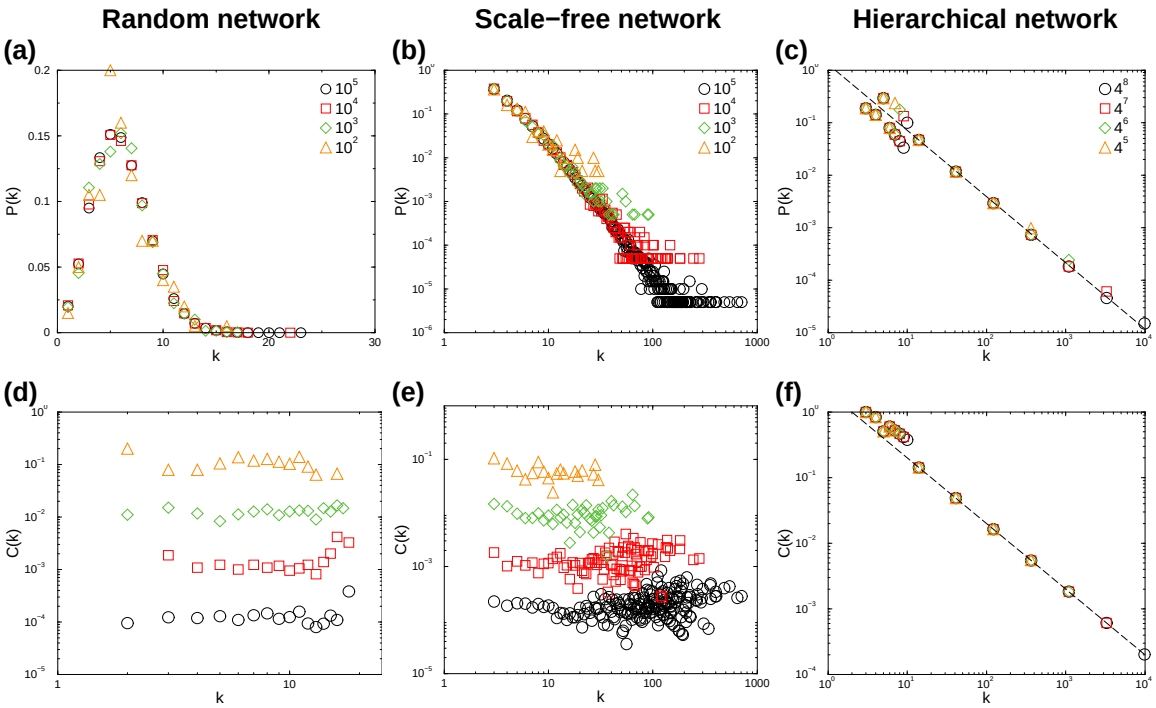
Random network
Scale−free network
Hierarchical network
(a)
(b)
(c)
(d)
(e)
(f)
P(k)
C(k)
k
10^5
10^4
10^3
10^2
4^8
4^7
4^6
4^5

Figure 4

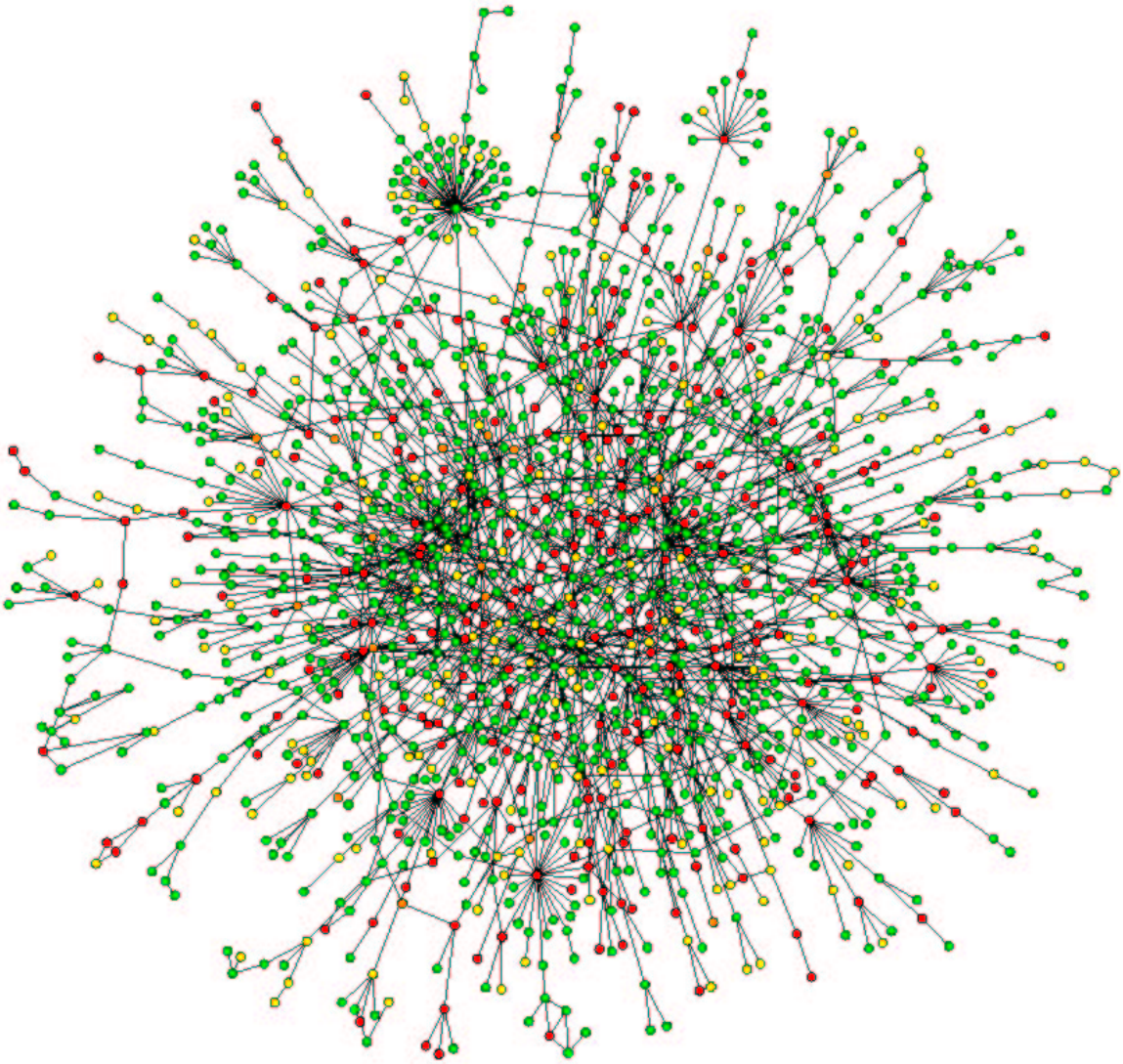

Figure 5

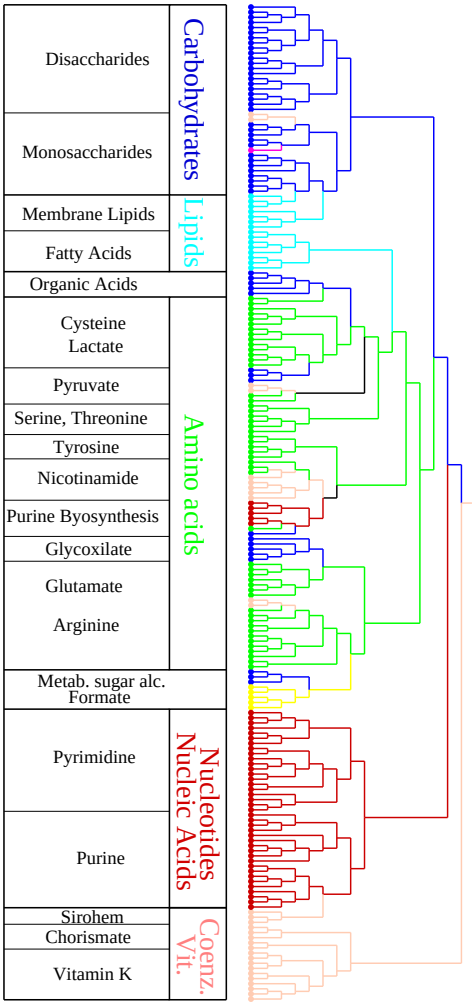
Carbohydrates
Lipids
Amino acids
Nucleotides
Nucleic Acids
Coenz.
Vit.
Disaccharides
Monosaccharides
Membrane Lipids
Fatty Acids
Organic Acids
Cysteine
Lactate
Pyruvate
Serine, Threonine
Tyrosine
Nicotinamide
Purine Byosynthesis
Glycoxilate
Glutamate
Arginine
Metab. sugar alc.
Formate
Pyrimidine
Purine
Sirohem
Chorismate
Vitamin K

Figure 6

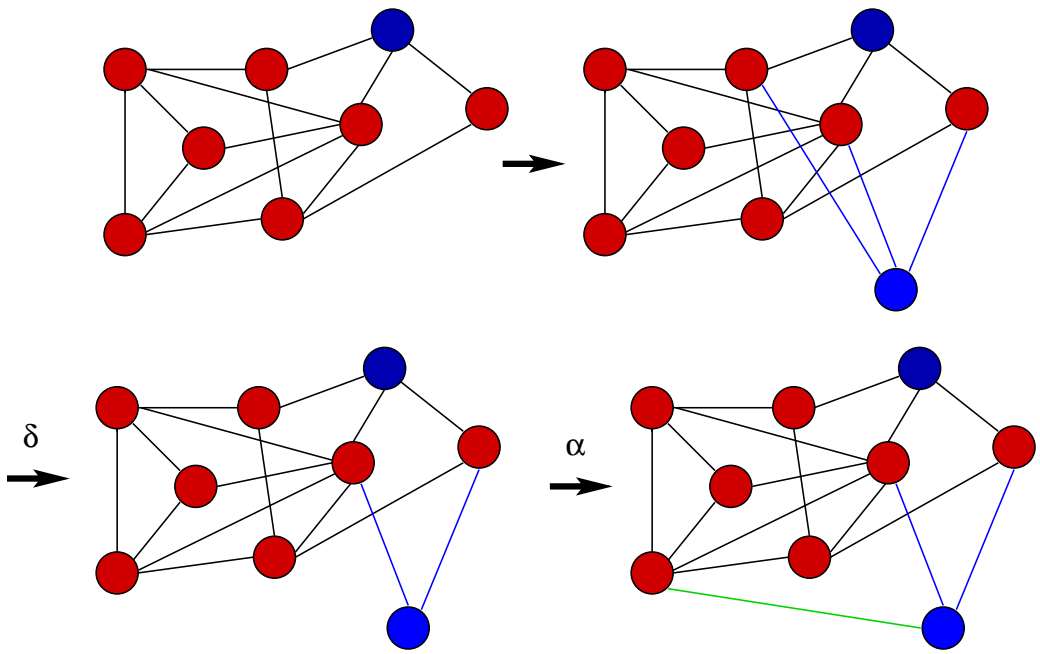

Figure 7

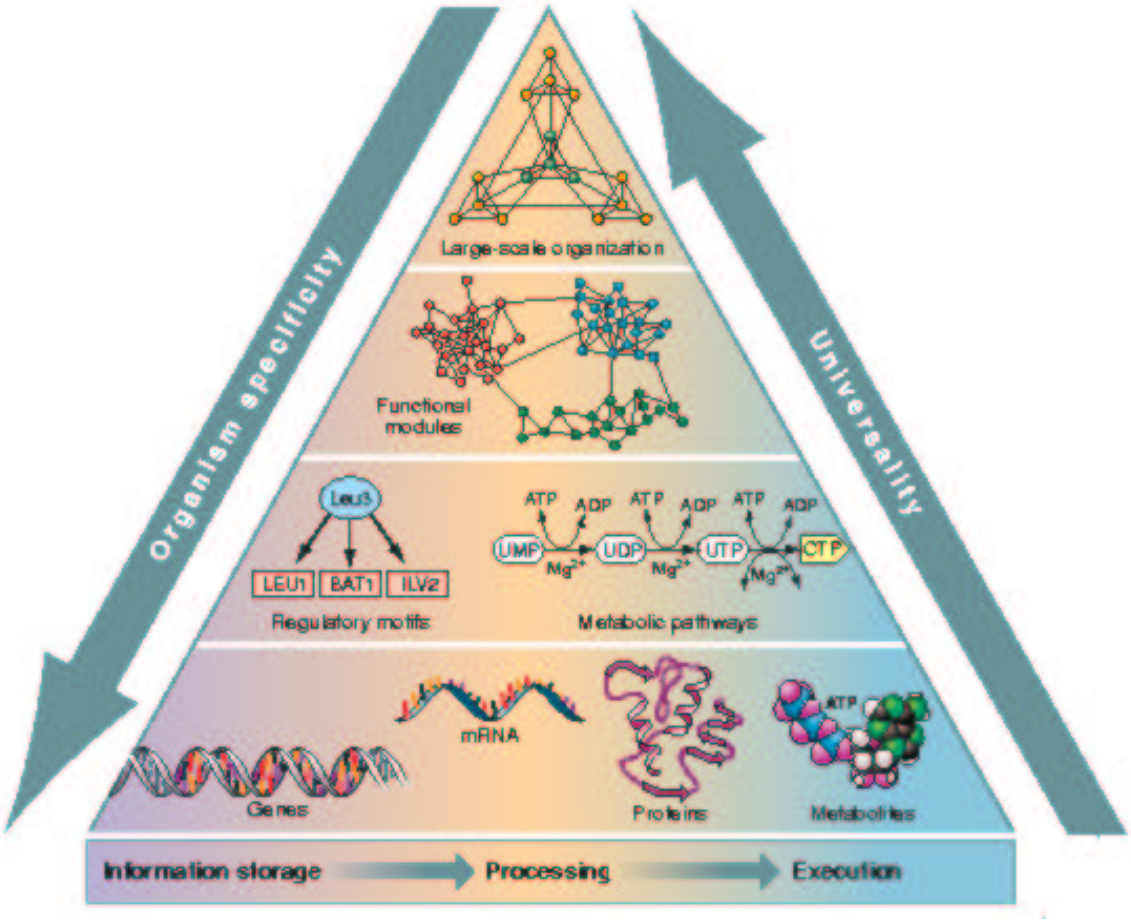
Organism specificity
Universality
Large-scale organization
Functional modules
Leu3
LEU1
BAT1
ILV2
Regulatory motifs
ATP
ADP
UMP
UDP
UTP
CTP
Metabolic pathways
mRNA
Genes
Proteins
Metabolites
Information storage
Processing
Execution

Figure 8