The global organization of cellular networks

Hawoong Jeong, Albert-László Barabási
Department of Physics
University of Notre Dame
Notre Dame, IN 46556, U.S.A.

Bálint Tombor, Zoltán N. Oltvai
Department of Pathology
Northwestern University Medical School
Chicago, IL 60611, U.S.A.

Abstract

Biologic phenomena arise as a sum of various cellular constituents and reactions seamlessly integrated into a complex functional network. Although the large-scale topology of metabolic networks has been established, the global relationship of all cellular constituents remains unknown. Here we show that the complete biochemical reaction network of 43 organisms from life's three domains, and separately their information pathway components, mirror the topologic scaling properties of their metabolic networks and display striking similarity to the organization of robust and error-tolerant scale-free networks. This may indicate that the global topology of cellular networks is organized identically in all species, and its form may represent a design pattern inherent to cellular life.

1 Introduction

Cells and microorganisms are complex systems in the natural world, which contain internal description of their structure, function, development and evolution encoded in the DNA sequences of their genes. As in complex systems in general, their phenotypic functions, such as chemotaxis, doesn't

solely arise from the biochemical or structural properties of their individual components, but also as a result of collective properties of the system as a whole. Consequently, a key emerging question in biology is how the complex biomolecular networks of various cellular constituents are organized to function as robust and dynamical systems.

Our ability to address in quantitative terms the structure of these cellular networks, has benefited significantly from recent advances in understanding the generic properties of complex networks. For example, until recently, complex networks have been modeled using the classical random network theory introduced by Erdös and Rényi (ER) [1]. This model assumes that each pair of nodes (that is, constituents) in the network is connected randomly with probability ρ, leading to a statistically homogeneous network, in which, despite the fundamental randomness of the model, most nodes have the same number of links, $\langle k \rangle$ (Fig.1a). In particular, the connectivity follows a Poisson distribution strongly peaked at $\langle k \rangle$ (Fig.1b), implying that the probability to find a highly connected node decays exponentially (i.e. $P(k) \sim \exp(-k)$ for $k \gg \langle k \rangle$). On the other hand, empirical studies on the structure of ecological webs, the Internet, and social networks have reported serious deviations from this random structure, demonstrating that these systems are described by scale-free networks [2] (Fig.1c), for which $P(k)$ follows a power-law, i.e. $P(k) \sim k^{-\gamma}$ (Fig. 1d). Unlike exponential networks, scale-free networks are extremely heterogeneous, their topology being dominated by a few highly connected nodes (hubs) which link the rest of the less connected nodes to the system (Fig. 1c). This topology makes scale-free networks highly tolerant against random errors and failures [4].

To begin to address the system-wide organization of cellular networks, we have recently determined the large-scale topologic organization of metabolic networks in 43 organisms from life's three domains [3]. Here we extend our analysis and show that the sum of all biochemical reactions and the ensuing networks within the same 43 organisms, and separately their information pathway components, mirror the topologic scaling properties of their metabolic domain and display striking similarity to the organization of robust and error-tolerant scale-free networks.

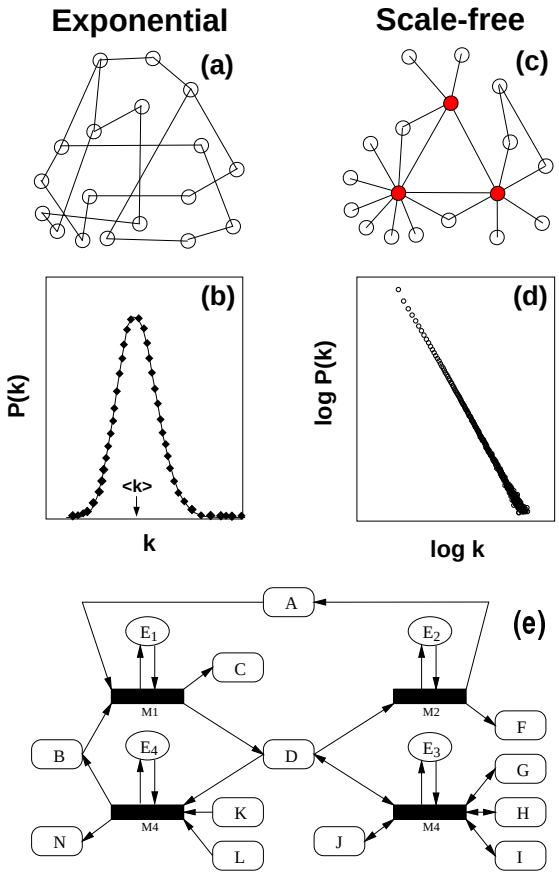

Figure 1: Attributes of fundamental network structures. (a) Representative structure of the network generated by the ER network model. (b) The network connectivity can be characterized by probability, $P(k)$, that a node has k links. For a random network $P(k)$ peaks strongly at $k = \langle k \rangle$ and decays exponentially for large k (that is, $P(k) \sim \exp(-k)$ for $k \gg \langle k \rangle$ and $k \ll \langle k \rangle$). (c) In the scale-free network most nodes have only a few links, but a few nodes, called hubs, have a very large number of links. (d) $P(k)$ for a scale-free network has no well-defined peak, and for large k, it decays as a power-law, $P(k) \sim k^{-\gamma}$, appearing as a straight line with slope $-\gamma$ on a log-log plot. (e) A portion of a hypothetical biochemical reaction network. Each substrate can be represented as a node of the graph, linked to one another through temporary substrate-enzyme complexes (black boxes) from which the products emerge as new nodes (substrates). The enzymes, which provide the catalytic scaffolds for the reactions, are circled.

2 RESULTS

2.1 Connectivity distribution of cellular networks

To address the large-scale structural organization of total biochemical reaction networks, we have examined the topologic properties of all pathways found in 43 different organisms from all three domains of life based on data deposited in the WIT database [5]. As we show in Fig. 1e, we have first established a graph theoretic representation of the biochemical reactions taking place in a given biochemical reaction network. In this representation, a biochemical reaction network is comprised of nodes, the various types of substrates (Fig 1e, A,B,C,...), that are connected to one another through links, which are the actual biochemical reactions. The physical entity of the link is the temporary substate-enzyme complex itself, in which enzymes provide the catalytic scaffolds for the reactions yielding products, which in turn can become educts for subsequent reactions.

Our first goal was to investigate the connectivity distribution of biochemical reaction networks as a whole by creating a single network that includes all pathways deposited in the WIT database for each organism, and to establish if their topology is best described by the inherently random and uniform exponential model [1] (Fig. 1a and b), or the highly heterogeneous scale-free model [2] (Fig. 1c and d). As illustrated in Fig. 2 (left column), our results convincingly indicate that the probability that a given substrate participates in k reactions follows a power-law distribution; in other words, biochemical reaction networks as a whole belong to the class of scale-free networks. As under physiological conditions a large number of biochemical reactions (links) in such a network are preferentially catalyzed in one direction (the links are directed), for each node we distinguish between incoming and outgoing links (Fig. 1e). For instance, in *Escherichia coli* the probability that a substrate participates as an educt in k biochemical reactions follows $P(k) \sim k^{-\gamma_{in}}$, with $\gamma_{in} = 2.1$, and the probability that a given substrate is produced by k different biochemical reactions follows a similar distribution, with $\gamma_{out} = 2.2$ (Fig. 2d). We find that a scale-free topology characterizes the total biochemical reaction networks in all organisms in all three domains of life (Fig. 2a,d,g) indicating the generic nature of this structural organization (Fig. 2j). Note that the power-law connectivity distribution has been independently found by Wagner and Fell for the *E. coli* metabolism [6].

The WIT database compartmentalizes cellular functions along bioengi-

neering principles and contains separate datasets for (1) intermediary metabolism and bioenergetics (core metabolism), (2) information pathways, (3) electron transport, (4) transmembrane transport, (5) signal transduction, and (6) structure and function of the cell, respectively. Due to current limitations in annotating certain segments of cellular function the core metabolic compartment represents the single largest portion of the database, followed by data deposited on information pathways. (The other portions of the database are substantially less developed, and are not amenable to statistical analysis with the mathematical tools currently available). As the topology of metabolic networks in the same 43 organisms was previously identified as scale-free [3], the question emerges if the above described topology of full biochemical reaction networks reflects on all their components, or is skewed due to overrepresentation of metabolic pathways in the WIT database. To clarify this issue we have separately investigated the topology of the information transfer components of biochemical reaction networks. Due to their small size, no statistical analysis could be performed on the information pathways of *A. pernix*, *E. nidulans*, *O. sativa* and *A. thaliana*. Yet, analysis on the remaining 39 organisms demonstrate that, similarly to that seen in metabolic networks (Fig. 2c,f,i,l), the probability that a given substrate participates in k information transfer reactions follows a power-law distribution both for incoming and outgoing links (Fig 2b,e,h,k). This indicates that the topology of full biochemical reaction networks apparently arises as a sum of a similar topologic organization of its major subcomponents.

2.2 The diameter of cellular networks is fixed

A general feature of many complex networks is their small-world character [7], meaning that any two nodes in the system can be connected by relatively short paths along existing links. In biochemical reaction networks these paths correspond to the biochemical pathway connecting two substrates. The degree of interconnectivity can be characterized by the network diameter, defined as the shortest biochemical pathway averaged over all pairs of substrates. For all non-biological networks examined to date the average connectivity of a node is fixed, which implies that the diameter of a network increases logarithmically with the addition of new nodes [2, 7, 8]. For biochemical reaction networks this implies that a more complex bacterium with higher number of enzymes and substrates, such as *E. coli*, would have a larger diameter than a simpler bacterium, such as *Mycoplasma genitalium*.

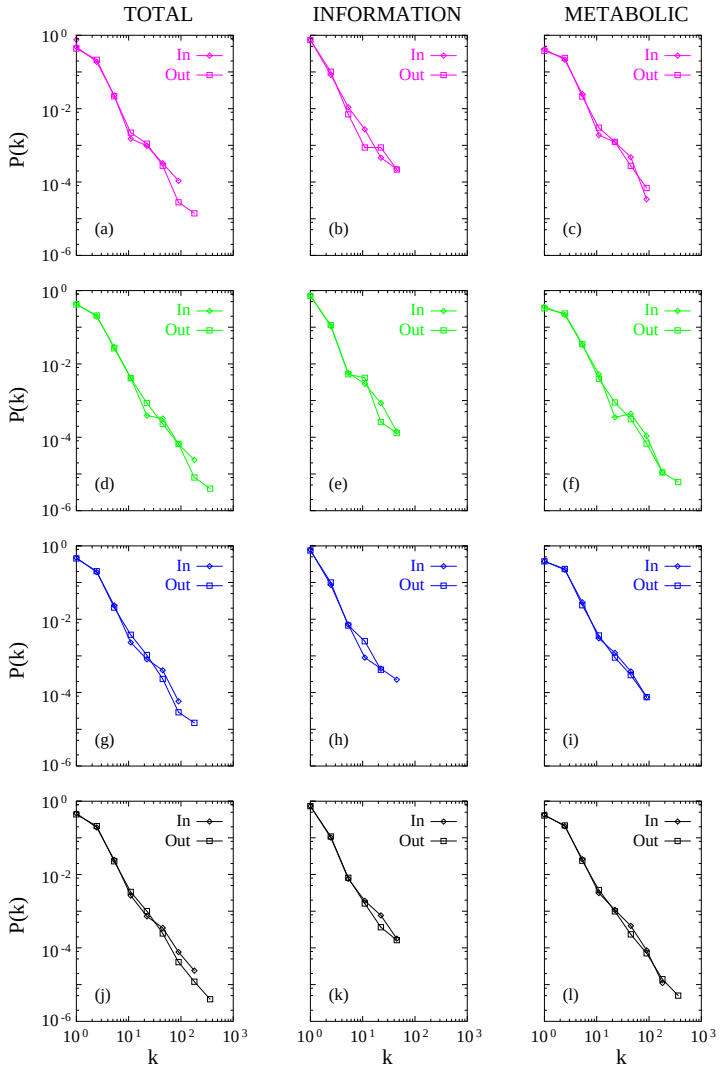

Figure 2: Connectivity distribution $P(k)$ for the substrates of (a-c) *A. fulgidus* (Archae), (d-f) *E. coli* (Bacterium), and (g-i) *C. elegans* (Eukaryote), shown on a log-log plot, counting separately the incoming (IN) and outgoing links (OUT) for each substrate, $k_{in}(k_{out})$ corresponding to the number of reactions in which a substrate participates as a product (educt). (j-l) The connectivity distribution averaged over all 43 organisms. The left column represents the complete biochemical reaction network (a,d,g,j) while the middle and right columns denote Information- (b,e,h,k) and Metabolic pathways (c,f,i,l), respectively.

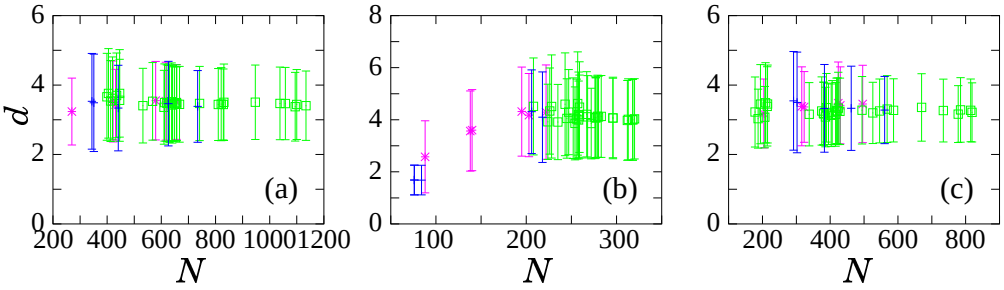

Figure 3: The average path length (diameter) for each of the 43 investigated organisms. The error bars correspond to the standard deviation. (a) complete- (b) information-, and (c) metabolic networks for Archaea [*], Bacteria [Ξ], and Eukaryotes [+], are shown. The horizontal axis (N) in (a-c) denotes the number of nodes in each organisms.

In contrast, we find that the diameter of biochemical reaction networks is the same for all 43 organisms, irrespective of the number of substrates found in the given species (Fig. 3a). This is in agreement to that found in the information- (Fig. 3b) and metabolic- (Fig. 3c) [3] portions of the network, and is consistent with a scenario in which with increasing organism complexity individual substrates on average are increasingly connected in order to maintain a relatively constant network diameter. Indeed, we find that for biochemical reaction networks as a whole the average number of reactions in which a certain substrate participate increases with the number of substrates found within the given organism (Fig. 4a,d). A similar trend is obvious for substrates of information- and metabolic pathways (Fig. 4b,c,e,f), the exception being incoming links for information pathway substrates for which no obvious trend is apparent (Fig. 4b).

2.3 Rank ordering the substrates

As the large-scale architecture of biochemical reaction networks rests on the most highly connected substrates, we need to investigate whether the same substrates act as hubs in all organisms, or if there are organism-specific differences in the identity of the most connected substrates. When we rank all the substrates in a given organism based on the number of links they have, we find that the ranking of the most connected substrates is practically identical for all 43 organisms. Also, only $\sim 5\%$ of all substrates that are found

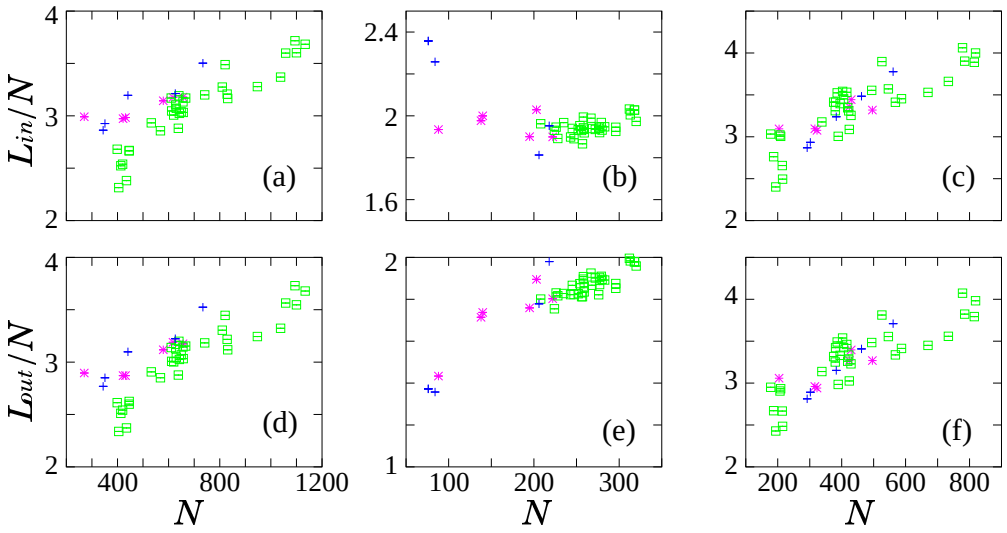

Figure 4: The average number of incoming (a-c) and outgoing (d-f) links per nodes for each organism. (a,d) complete-, (b,e) information-, and (c,f) metabolic networks for Archaea [*], Bacteria [Ξ], and Eukaryotes [+], are shown. The horizontal axis (N) in (a-f) denotes the number of nodes in each organisms.

in all 43 organisms are present in all species. These substrates represent the most highly connected substrates found in any individual organism, indicating the generic utilization of the same substrates by each species. In contrast, species-specific differences among various organisms emerge for less connected substrates. To quantify this observation, we examined the standard deviation (σ_r) of the rank for substrates that are present in all 43 organisms. As shown in Fig. 5a, we find that sr increases with the average rank order, $\langle r \rangle$, implying that the most connected substrates have a relatively fixed position in the rank order, but the ranking of less connected substrates is increasingly species-specific. A separate examination of information pathway components (Fig. 5b) confirms these results and is in agreement to that seen in metabolic networks (Fig. 5c) [3]. Thus, the large-scale structure of biochemical reaction networks is identical for all 43 species, being dominated by the same highly connected substrates, while less connected substrates preferentially serve as educt or product of species-specific enzymatic activities.

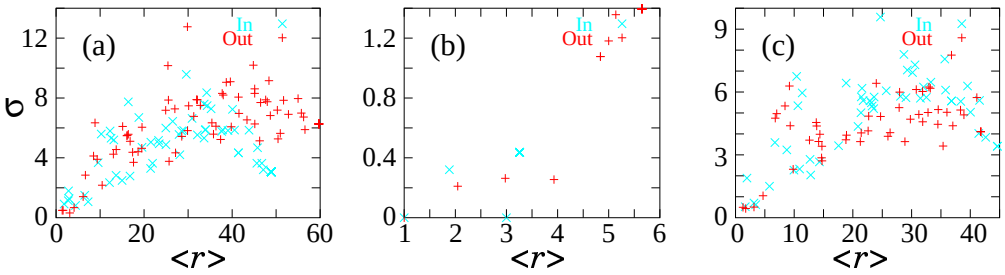

Figure 5: Standard deviation of the substrate ranking (σ_r) in the (a) complete-, (b) information-, and (c) metabolic network for Archaea, Bacteria, and Eukaryotes, as a function of the average ranking, $\langle r \rangle_o$ for substrates present in all 43 investigated organisms.

3 DISCUSSION

Biochemical reaction networks, coherently first identified and comprehensively described by integrated pathway-genome databases [9], represent the possibilities that are encoded within a given genome. Of note, in explicit form genomes in themselves contain information only for the genes and their order within the genome. Yet, the elementary units of these networks, which are the various biochemical reactions, exist because (genome-encoded) enzymes carry out defined reactions at specific rates. Thus, the higher-level coherence of their topology is implicitly generated by the properties of the products of the genetic elements (i.e., enzymes), and the system presumably exists as is, as a result of surviving the test of natural selection in the context of the particular ecological niche the specific organism occupies.

But what is the evolutionary origin of the scale-free structure and the associated power law distribution observed in the topology of all studied organisms? Although at present no definitive answer can be provided, the scale-free structure in complex generic networks has a dynamical origin, requiring the coexistence of growth and preferential attachment [2]. This implies that cellular networks emerged starting from a small number of simple elements, subsequent reactions generating new compounds (nodes) that were able to react with those substrates already present in the system. This scenario, for example, is in agreement with current theories concerning the chemical origin of intermediary metabolism, in which a prebiotic nonenzymatic reductive citric acid cycle with network autocatalytic properties served as basis for the subsequent synthesis of more complex elements [10, 11]. As preferential

attachment predicts that the first metabolic substrates are also the most connected ones, the finding that in *E. coli* the majority of the most connected metabolites coincide with those that presumably first emerged during the assembly of cell metabolism [6] further supports this concept.

Finally, the uncovered organizational principles may represent deeper patterns common to life that are not necessarily restricted to terrestial biology [11]. For instance, at least theoretically, life has no fundamental requirement for purines and pyrimidines as units of information storage and transfer in the form of DNA and RNA. Yet, biochemical reactions, such as those constituting cellular metabolism, are likely to represent a universal attribute of all living systems. The uniform cellular network topology observed in all 43 organisms strongly suggests that, irrespective of their individual building blocks or species-specific reaction pathways, natural selection favors a robust and error tolerant scale-free architecture that appears to provide an optimal structural organization of cellular existence.

[The *Methods and Procedures* pertaining to this paper can be found at http://www.nd.edu/~networks/cell.]

References

[1] Bollobas, B. Random Graphs, Academic Press, London, 1985.

[2] Barabasi, A.-L. and Albert, R. Emergence of scaling in random networks., Science. **286**: 509-12, 1999.

[3] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A.-L. The large-scale organization of metabolic networks, Nature. **407**: 651-654, 2000.

[4] Albert, R., Jeong, H. and Barabási, A.-L. Nature **406**: 378-382, 2000.

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

[6] Fell, D.A. and Wagner, A. in Animating the cellular map (eds Hofmeyr, J.-H. S., Rohwer, J. M. and Snoep, J.L.) 79-85 (Stellenbosch University Press, Stellenbosch, 2000).

[7] Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world' networks, Nature. **393**: 440-2, 1998.

[8] Barthelemy, M. and Amaral, L. A. N. Small-world networks: Evidence for a crossover picture, Phys Rev Lett. **82**: 3180, 1999.

[9] Karp, P. D., Krummenacker, M., Paley, S., and Wagg, J. Integrated pathway-genome databases and their role in drug discovery, Trends Biotechnol. **17**: 275-81, 1999.

[10] Wachtershauser, G. The origin of life and its methodological challenge., J. Theor. Biol. **187**: 483-694, 1997.

[11] Morowitz, H. J., Kostelnik, J. D., Yang, J., and Cody, G. D. The origin of intermediary metabolism, Proc Natl Acad Sci U S A. **97**: 7704-8, 2000.