The large-scale organization of metabolic networks

H. Jeong[1], B. Tombor[2], R. Albert[1], Z. N. Oltvai[2] and A.-L. Barabási[1]

[1]Department of Physics, University of Notre Dame, Notre Dame, IN 46556, and

[2]Department of Pathology, Northwestern University Medical School, Chicago, IL 60611

In a cell or microorganism the processes that generate mass, energy, information transfer, and cell fate specification are seamlessly integrated through a complex network of various cellular constituents and reactions [1]. However, despite the key role these networks play in sustaining various cellular functions, their large-scale structure is essentially unknown. Here we present the first systematic comparative mathematical analysis of the metabolic networks of 43 organisms representing all three domains of life. We show that, despite significant variances in their individual constituents and pathways, these metabolic networks display the same topologic scaling properties demonstrating striking similarities to the inherent organization of complex non-biological systems [2]. This suggests that the metabolic organization is not only identical for all living organisms, but complies with the design principles of robust and error-tolerant networks [2-5], and may represent a common blueprint for the large-scale organization of interactions among all cellular constituents.

An important goal in biology is to uncover the fundamental design principles that provide the common underlying structure and function in all cells and microorganisms [6-13]. For example, it is increasingly appreciated that the robustness of various cellular processes seen in a cell or microorganism is rooted in the dynamic interactions among its many constituents [14, 15], such as proteins, DNA, RNA, and small molecules. Yet, the large-scale design principles that integrate these interactions into a complex system are poorly understood [1]. Recent scientific developments, however, significantly improve our ability to identify such principles. Large-scale genome sequencing projects have provided complete sequence information for more than two dozen prokaryotes and a few

eukaryotes, allowing the identification of known pathways in the annotated genome of an organism [16]. This, combined with firmly established data from the biochemical literature, led to the development of integrated pathway-genome databases [17-19] that provide organism-specific connectivity maps of metabolic- and, to a lesser extent, various other cellular networks. Yet, due to the large number and the diversity of the constituents and reactions forming such networks, these maps are extremely complex, offering only limited insight into the design principles governing the organization of these systems. Our ability to address in quantitative terms the structure of these cellular networks, however, has benefited from recent advances in understanding the generic properties of complex networks [2].

Until recently, complex networks have been modeled using the classical random network theory introduced by Erdös and Rényi (ER) [20, 21]. The ER model assumes that each pair of nodes (i.e., constituents) in the network is connected randomly with probability p, 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. $P(k) \sim e^{-k}$ for $k \gg \langle k \rangle$). On the other hand, recent empirical studies on the structure of the World-Wide Web [22], Internet [23], and social networks [2] 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). Since the distinction between the scale-free and exponential networks emerges as a result of simple dynamical principles [24, 25], understanding the large-scale structure of cellular networks can provide not only valuable and perhaps universal structural information, but could also lead to a better understanding of the dynamical processes that generated these networks. In this respect the emergence of power law distribution is intimately linked to the growth of the network in which new nodes are preferentially

attached to already established nodes [2], a property that is also thought to characterize the evolution of biological systems [1].

To address the large-scale structural organization of cellular networks, we have examined the topologic properties of the core metabolic network of 43 different organisms based on data deposited in the WIT database [18]. This integrated pathway-genome database predicts the existence of a given metabolic pathway primarily based on the annotated genome of an organism, i.e. on the presence of the open reading frame (ORF) of an enzyme that catalyzes a given metabolic reaction. As 18 of the 43 organisms deposited in the database are not yet fully sequenced, and a substantial portion of the identified ORFs are functionally unassigned, the list of enzymes, and consequently the list of substrates and reactions (Table 1), will certainly be expanded in the future. Nevertheless, this publicly available database represents our current best approximation for the metabolic pathways in 43 organisms and provides sufficient data for their unambiguous statistical analysis (see Methods and Supplementary Material [26]).

As we illustrate in Fig. 1e, we have first established a graph theoretic representation of the biochemical reactions taking place in a given metabolic network. In this representation, a metabolic network is built up of nodes, which are the substrates that are connected to one another through links, which are the actual metabolic reactions. The physical entity of the link is the temporary educt-educt complex itself, in which enzymes provide the catalytic scaffolds for the reactions yielding products, which in turn can become educts for subsequent reactions. Of note, a few reactions are spontaneous and require no enzymatic activity (Fig. 1e, lower left corner). This representation allows us to systematically investigate and quantify the topologic properties of various metabolic networks using the tools of graph theory and statistical mechanics [21]. Our first goal was to identify the structure of the metabolic networks, i.e., to establish if their topology is best described by the inherently random and uniform exponential model [21] (Fig. 1a and b), or the highly heterogeneous scale-free model [2] (Fig. 1c and d). As illustrated in Fig. 2, our results convincingly indicate that the probability that a given

substrate participates in *k* reactions follows a power-law distribution, i.e., metabolic networks belong to the class of scale-free networks. Since under physiological conditions a large number of biochemical reactions (links) in a metabolic network are preferentially catalyzed in one direction (i.e. the links are directed), for each node we distinguish between incoming and outgoing links (Fig. 1e). For instance, in *E. coli* the probability that a substrate participates as an educt in *k* metabolic reactions follows $P(k) \sim k^{-\gamma_{in}}$, with $\gamma_{in} = 2.2$, and the probability that a given substrate is produced by *k* different metabolic reactions follows a similar distribution, with $\gamma_{out} = 2.2$ (Fig. 2b). We find that scale-free networks describe the metabolic networks in all organisms in all three domains of life (Fig. 2a-c) [26], indicating the generic nature of this structural organization (Fig. 2d).

A general feature of many complex networks is their small-world character [27], i.e. any two nodes in the system can be connected by relatively short paths along existing links. In metabolic networks these paths correspond to the biochemical pathway connecting two substrates (Fig. 3a). The degree of interconnectivity of a metabolic network 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, 27, 28]. For metabolic 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 *M. genitalium*. In contrast, we find that the diameter of the metabolic network is the same for all 43 organisms, irrespective of the number of substrates found in the given species (Fig. 3b). This is surprising and unprecedented, and is possible only if with increasing organism complexity individual substrates are increasingly connected in order to maintain a relatively constant metabolic network diameter (Fig. 3a inset). Indeed, we find that the average number of reactions in which a certain substrate participates increases with the number of substrates found within the given organism (Fig. 3c and d).

An important consequence of the power-law connectivity distribution is that a few hubs dominate the overall connectivity of the network (Fig. 1c), and upon the sequential removal of the most-connected nodes the diameter of the network rises sharply, the network eventually disintegrating into isolated clusters that are no longer functional. Yet, scale-free networks also demonstrate unexpected robustness against random errors [5]. To examine if metabolic networks display a similar error tolerance we performed computer simulations on the metabolic network of the bacterium, *E. coli*. Upon removal of the most connected substrates the diameter increases rapidly, illustrating the special role these metabolites play in maintaining a constant metabolic network diameter (Fig. 3e). However, when randomly chosen *M* substrates were removed, -mimicking the consequence of random mutations of catalyzing enzymes-, the average distance between the remaining nodes was not affected, indicating a striking insensitivity to random errors. Indeed, *in-silico* and *in-vivo* mutagenesis studies indicate a remarkable fault tolerance upon removal of a substantial number of metabolic enzymes from the *E. coli* metabolic network [29]. Of note, data similar to that shown in Fig. 3e have been obtained for all investigated organisms, without detectable correlations with their evolutionary position.

Since the large-scale architecture of the metabolic network rests on the most highly connected substrates, we need to address whether the same substrates act as hubs in all organisms, or if there are major organism-specific differences in the identity of the most connected substrates. When we rank order all the substrates in a given organism based on the number of links they have (Table 1), we find that the ranking of the most connected substrates is practically identical for all 43 organisms. Also, only ~4% of all substrates that are found 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. 3f, we find that σ_r 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. Thus, the large-scale structure of the metabolic network 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.

The contemporary topology of a metabolic network reflects a long evolutionary process molded in general for a robust response towards internal defects and environmental fluctuations and in particular to the ecological niche the specific organism occupies. As a result, one expects that these networks are far from being random, and our data demonstrate that the large-scale structural organization of metabolic networks is indeed highly similar to that of robust and error-tolerant networks [2, 5]. The uniform network topology observed in all 43 organisms strongly suggests that, irrespective of their individual building blocks or species-specific reaction pathways, the large-scale structure of metabolic networks is identical in all living organisms.

A unique feature of metabolic networks, as opposed to that seen in non-biological scale-free networks, is the apparent conservation of the network diameter in all living organisms. Within the special characteristics of living systems this attribute may represent an additional survival and growth advantage, since a larger diameter would attenuate the organism's ability to efficiently respond to external changes or internal errors. For example, should the concentration of a substrate suddenly diminish due to mutation in its main catalyzing enzyme, offsetting the changes would involve the activation of longer alternative biochemical pathways, and consequently the synthesis of more new enzymes, than within a smaller metabolic network diameter.

But how generic these principles are for other cellular networks (e.g., information transfer, cell cycle)? While the current mathematical tools do not allow unambiguous statistical analysis of the topology of other networks due to their relatively small size, our preliminary analysis suggest that connectivity distribution of non-metabolic pathways also follows a power-law distribution, indicating that cellular networks as a whole are scale-free networks. Therefore, the evolutionary selection of a robust

and error tolerant architecture may characterize all cellular networks, for which the scale-free topology with a conserved network diameter appears to provide an optimal structural organization.

Methods

Database preparation: For our analyses of core cellular metabolisms we used the "Intermediate metabolism and Bioenergetics" portions of the WIT database [18] (http://igweb.integratedgenomics.com/IGwit/), that predicts the existence of a metabolic pathway in an organism primarily based on the annotated genome of the organism, i.e., on the presence of the open reading frame (ORF) of an enzyme that catalyzes a given metabolic reaction. As of December 1999, this database provides description for 6 archaea, 32 bacteria and 5 eukaryota. The downloaded data were manually rechecked, removing synonyms and substrates without defined chemical identity.

Construction of metabolic network matrices: Biochemical reactions described within a WIT database are composed of substrates and enzymes connected by directed links. For each reaction, educts and products were considered as nodes connected to the temporary educt-educt complexes and associated enzymes. Bi-directional reactions were considered separately. For a given organism with N substrates, E enzymes and R intermediate complexes the full stochiometric interactions were compiled into an $(N+E+R) \times (N+E+R)$ matrix, generated separately for each of the 43 organisms.

Connectivity distribution [$P(k)$]: Substrates generated by a biochemical reaction are products, and are characterized by incoming links pointing to them. For each substrate we have determined k_{in}, and prepared a histogram for each organism, providing how many substrates have exactly k_{in} =0,1,…. Dividing each point of the histogram with the total number of substrates in the organism provided $P(k_{in})$, or the probability that a substrate has k_{in} incoming links. Substrates that participate as educts in a reaction have outgoing links. We have performed the analysis described above for k_{in}, determining the number of outgoing links (k_{out}) for each substrate. To reduce noise logarithmic binning was applied.

Biochemical pathway lengths [$\Pi(l)$]: For all pairs of substrates, the shortest biochemical pathway, $\Pi(l)$ (i.e., the smallest number of reactions by which one can reach substrate B from substrate A) were determined using a burning algorithm. From $\Pi(l)$ we determined the diameter, $D = \sum_l l \cdot \Pi(l) \Big/ \sum_l \Pi(l)$, which represents the average path length between any two substrates.

Substrate ranking [$\langle r \rangle_o$, $\sigma(r)$]: Substrates present in all 43 organisms (i.e., a total of 51 substrates) were ranked based on the number of links each had in each organisms, having considered incoming and outgoing links separately (r =1 were assigned for the substrate with the largest number of connections, and r =2 for second most connected one, etc.). This way for each substrate a well-defined r value in each organism was defined. The average rank $\langle r \rangle_o$ for each substrate was determined by averaging r over the 43 organisms. We also determined the standard deviation, $\sigma(r) = \langle r^2 \rangle_o - \langle r \rangle_o^2$ for all 51 substrates present in all organisms.

Analysis of the effect of database errors: Of the 43 organisms whose metabolic network we have analyzed the genome of 25 has been completely sequenced (5 Archae, 18 Bacteria, 2 Eukaryotes), while the remaining 18 are only partially sequenced. Therefore two major sources of possible errors in the database could affect our analysis: (a) the erroneous annotation of enzymes and consequently, biochemical reactions; for the organisms with completely sequenced genomes this is the likely source of error. (b) reactions and pathways missing from the database; for organisms with incompletely sequenced genomes both (a) and (b) are of potential source of error. We investigated the effect of database errors on the validity of our findings, the results being presented in the Supplementary Material [26], indicating that the results offered in this paper are robust to these errors.

Acknowledgement

We would like to acknowledge all members of the WIT project for making this extremely valuable database publicly available for the scientific community. We also thank C. Waltenbaugh and H. S. Seifert for comments on the manuscript. Research at the University of Notre Dame was supported by the National Science Foundation, and at Northwestern University by grants from the National Cancer Institute.

Correspondence and requests for materials should be addressed to A-L.B. (alb@nd.edu) or Z.N.O. (zno008@nwu.edu).

References

1. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47-52 (1999).

2. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509-12 (1999).

3. West, G.B., Brown, J.H. & Enquist, B.J. The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science* **284**, 1677-9 (1999).

4. Banavar, J.R., Maritan, A. & Rinaldo, A. Size and form in efficient transportation networks. *Nature* **399**, 130-2 (1999).

5. Albert, R., Jeong, H. & Barabási, A.-L. The Internet's Achilles' heel: Error and attack tolerance of complex networks. *Nature* , in press (2000).

6. Ingber, D.E. Cellular tensegrity: defining new rules of biological design that govern the cytoskeleton. *J Cell Sci* **104**, 613-27 (1993).

7. Bray, D. Protein molecules as computational elements in living cells. *Nature* **376**, 307-12 (1995).

8. McAdams, H.H. & Arkin, A. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet* **15**, 65-9 (1999).

9. Gardner, T.S., Cantor, C.R. & Collins, J.J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339-42 (2000).

10. Elowitz, M.B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335-8 (2000).

11. Hasty, J., Pradines, J., Dolnik, M. & Collins, J.J. Noise-based switches and amplifiers for gene expression. *Proc Natl Acad Sci U S A* **97**, 2075-2080 (2000).

12. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590-3 (2000).

13. Kirschner, M., Gerhart, J. & Mitchison, T. Molecular "vitalism". *Cell* **100**, 79-88 (2000).

14. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913-7 (1997).

15. Bhalla, U.S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science* **283**, 381-7 (1999).

16. Karp, P.D., Krummenacker, M., Paley, S. & Wagg, J. Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol* **17**, 275-81 (1999).

17. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

18. Overbeek, R.*, et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* **28**, 123-125 (2000).

19. Karp, P.D.*, et al.* The EcoCyc and MetaCyc databases. *Nucleic Acids Res* **28**, 56-59 (2000).

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

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

22. Albert, R., Jeong, H. & Barabási, A.-L. Diameter of the World-Wide Web. *Nature* **400**, 130-131 (1999).

23. Faloutsos, M., Faloutsos, P. & Faloutsos, C. On Power-Law Relationships of the Internet Topology. *ACM SIGCOMM 99, Comp. Comm. Rev.* **29**, 251 (1999).

24. Amaral, L.A.N., Scala, A., Barthelemy, M. & Stanley, H.E. Classes of behavior of small-world networks. *http://xxx.lanl.gov/abs/cond-mat/0001458* (2000).

25. Dorogovtsev, S.N. & Mendes, J.F.F. Evolution of reference networks with aging. *http://xxx.lanl.gov/abs/cond-mat/0001419* (2000).

26. www.nd.edu/~networks/cell.

27. Watts, D.J. & Strogatz, S.H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440-2 (1998).

28. Barthelemy, M. & Amaral, L.A.N. Small-world networks: Evidence for a crossover picture. *Physical Review Letters* **82**, 3180 (1999).

29. Edwards, J.S. & Palsson, B.O. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A* **97**, 5528-33 (2000).

FIGURE LEGENDS

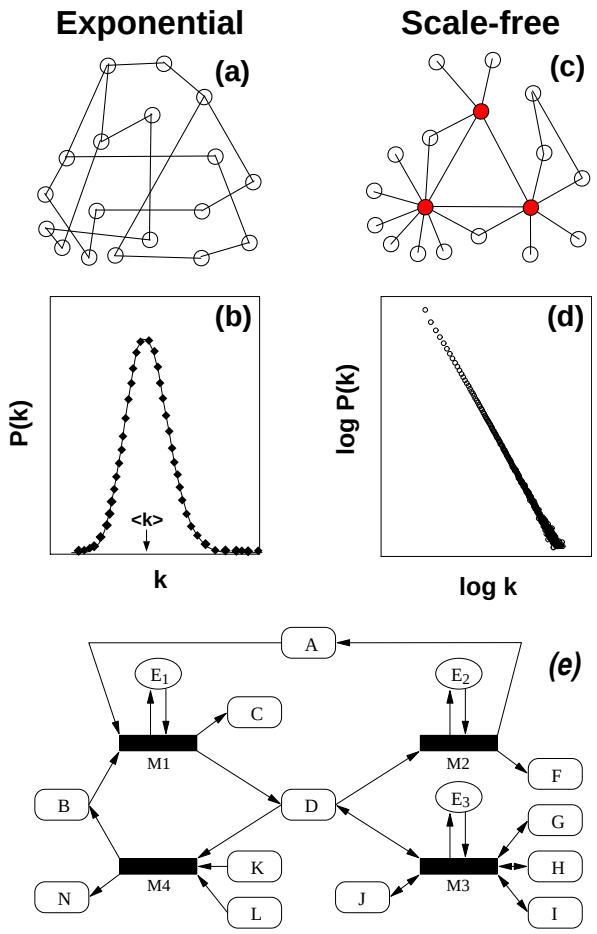

Figure 1

(**a**) The classical random network model of Erdös and Rényi assumes that each pair of nodes are connected randomly, with probability *p*. Since links are assigned randomly, nodes can have different number of links (connectivity). (**b**) The network connectivity can be characterized by the histogram of the number of nodes with *k* links, which, after normalization, gives the probability, *P(k),* that a node has *k* links. For a random network *P(k)* is strongly peaked at $k = \langle k \rangle$ and decays exponentially for large *k* (i.e. $P(k) \sim e^{-k}$ for $k \gg \langle k \rangle$ and $k \ll \langle k \rangle$), implying that the majority of nodes have the same connectivity $k \approx \langle k \rangle$. (**c**) A scale-free network has a drastically different topology: most nodes have only a few links, but a few nodes, called hubs (red), have a very large number of links, effectively connecting the rest of the nodes together into a fully connected network. (**d**) The connectivity distribution, *P(k)*, for a scale-free network has no well-defined peak, indicating the absence of a characteristic scale. For large *k*, *P(k)* decays as a power-law, $P(k) \sim k^{-\gamma}$, thus *P(k)* appears as a straight line with slope $-\gamma$ on a log-log plot. (**e**) The substrates and the reactions in a metabolic network can be uniquely described using a graph theoretic representation. Each substrate represents a node of the graph, denoted by A, B, C, …. These nodes are linked to one another through temporary educt-educt complexes (black boxes, M_1, M_2, …), from which the products emerge as new nodes (substrates). The enzymes that provide the catalytic scaffolds for the reactions are also shown (denoted by E_1, E_2, …). Since under *in vivo* conditions some reactions are reversible, while others are irreversible, in the graph we have both directed (simple arrow) or undirected (double arrow) links. The complexity of the metabolic network comes from the fact that some substrates can participate in multiple reactions, and can act both as educt or product of different reactions.

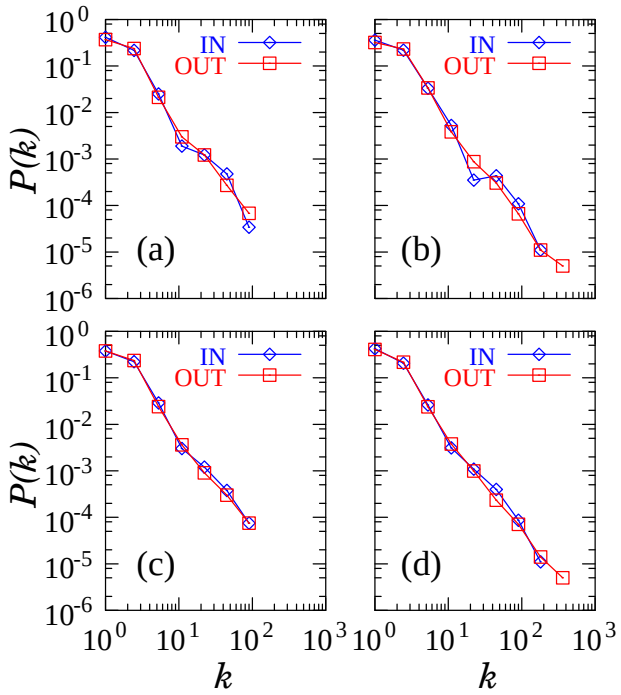

Figure 2

Connectivity distribution *P(k)* for the substrates in (**a**) *A. fulgidus* (Archae) (**b**) *E. coli* (Bacterium) (**c**) *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 an product (educt) (see Fig. 1**e**). Practically indistinguishable plots have been obtained for all 43 organisms investigated. The characteristics of the three organisms shown in **a**-**c** and the exponents γ_{in} and γ_{out} for all organisms are given in Table 1. (**d**) The connectivity distribution averaged over all 43 organisms. The exponent γ was determined from the slope of the plots, calculated using a least square fit method.

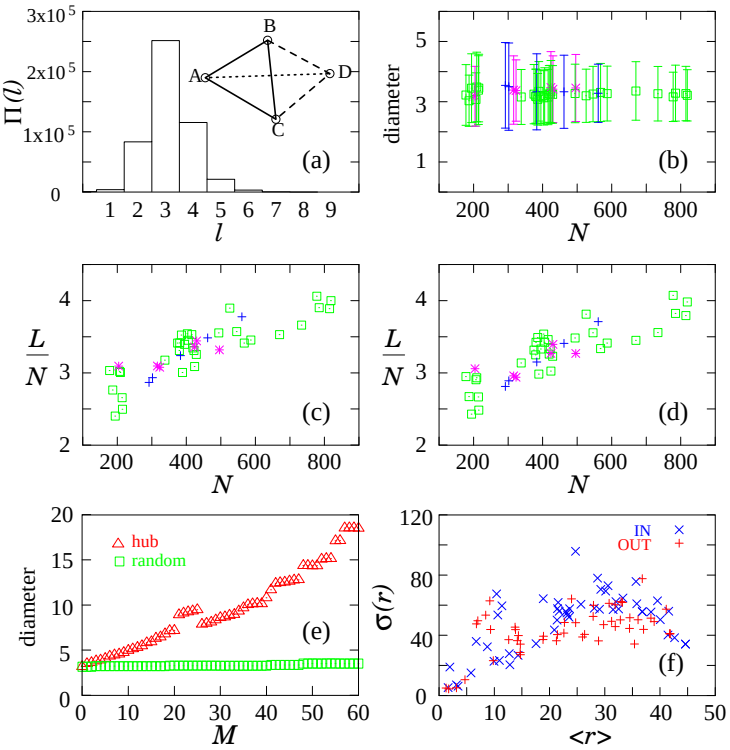

Figure 3

(**a**) The histogram of the biochemical pathway lengths, ℓ, in the bacterium, *E. coli*. For example in Fig. 1e the biochemical pathway distance between the substrates B and D is ℓ=1, since these can be linked through a single reaction (B $\xrightarrow{E_2}$ D), but between B and F we have ℓ=2, counted, for example, along the B $\xrightarrow{E_2}$ D $\xrightarrow{E_1}$ F pathway. The figure shows the histogram of the shortest biochemical pathway lengths for all pairs of substrates in *E. coli*. The average path length gives the diameter of the network. The strong peak of $\Pi(\ell)$ indicates that in *E. coli* most substrates are connected by a path of length ℓ=3, i.e. the average diameter of the network is close to three (the precise value is *D*=3.2). Inset: Schematic illustration of the changes in the network diameter during growth. In a simple network containing three nodes (A, B and C) each having 2 links (solid lines), the distance between any two nodes is equal to 1, thus the diameter of the network is 1. When a new node (D) with two links (BD and CD, dashed lines) is added to the system, the diameter increases to $D=(\ell_{AB}+\ell_{AC}+\ell_{BC}+\ell_{BD}+\ell_{DC}+\ell_{AD})/6 = 7/6$ since $\ell_{AD}=2$ and all other distances are ℓ=1. Adding an extra link (AD, dotted line) will decrease the diameter, bringing it back to one, while the average connectivity increases from 2 to 3, demonstrating that an unchanged diameter can be maintained through increasing the average connectivity. (**b**) The diameter for all 43 investigated organisms, indicating that the diameter is constant within the error bars. The error bars in (**b**) correspond to the standard deviation $\sigma \sim \langle \ell^2 \rangle - \langle \ell \rangle^2$ as determined from $\Pi(\ell)$ (shown in (**a**) for *E. coli*). The horizontal axis denotes the number of nodes in each organism. Archaea are shown in magenta, bacteria in green, and eukaryotes in blue, a color code used in Table 1 as well. (**c**) The average number of incoming links per node for each bacterium studied in the database, calculated by dividing the total number of incoming links (L_{in}) with the number of substrates (N) in a given organism. (**d**) The same as in (**c**) but for outgoing links. (**e**) The effect of substrate removal on the diameter of the E. coli bacteria. In the upper curve (Δ) we select and remove the nodes in an inverse order of connectivity, starting with the most connected substrate first (largest hub), and continuing in decreasing order of the number of links, indicating that the diameter is sensitive to the presence/absence of the most connected nodes. In the bottom curve ($\square$) nodes are selected and removed randomly. The unchanged diameter indicates the high degree of error tolerance of the metabolic network. Note that M=60 corresponds to 8% of the total number of substrates in *E. coli*. (**f**) Standard deviation of the substrate ranking (σ_r) as a function of the average ranking, $\langle r \rangle_o$ for substrates present in all 43 investigated organisms. For each organism we ranked all substrates based on the number of links they had, assigning r=1 for the most connected substrate, r=2 for next less connected one, and so on. We then determined $\langle r \rangle_o$ for a given substrate, where $\langle\ \rangle_o$ denotes averaging the ranking of a given substrate over all 43 organisms, and the standard deviation, $\sigma_r = \langle r^2 \rangle_o - \langle r \rangle_o^2$. A substrate with $\sigma_r = 0$ and average rank $\langle r \rangle_o$ has rank $r = \langle r \rangle_o$ in each of the 43 organisms, while an increasing σ_r signals increased variability in the ranking of the substrate. Note that only 51 substrates are present in all organisms, a mere 4% of the total, indicating high degree of differentiation between different species.

Table 1.

No.	Name	*N*	*L*(IN)	*L*(OUT)	*R*	*E*	γ_{in}	γ_{out}	*D*	Hub(IN)	Hub(OUT)
1	A. pernix	204	588	575	178	135	2.2	2.2	3.2	bacdelgfij	adbcegipfh
2	A. fulgidus	496	1527	1484	486	299	2.2	2.2	3.5	abcdghefjk	adbijchemf
3	M. thermoautotrophicum	430	1374	1331	428	280	2.2	2.2	3.4	abcdgefkhi	adbicejfkh
4	M. jannaschii	424	1317	1272	415	264	2.2	2.3	3.5	abcdgekfnh	adbciejkhf
5	P. furiosus	316	901	867	283	191	2.0	2.3	3.4	abcdgeknfh	dabceipjhf
6	P. horikoshii	323	914	882	288	196	2.0	2.2	3.4	abdcgefknl	dabciepjhq
7	A. aeolicus	419	1278	1249	401	285	2.1	2.2	3.3	bcadgefhki	adbciejgfh
8	C. pneumoniae	194	401	391	134	84	2.2	2.3	3.4	bdcagfleri	dabciergfp
9	C. trachomatis	215	479	462	158	94	2.2	2.4	3.5	bdacgfelrm	dbaciegrfp
10	Synechocystis sp.	546	1782	1746	570	370	2.0	2.2	3.3	abcdegfhjk	adbicjehfg
11	P. gingivalis	424	1192	1156	374	254	2.2	2.2	3.3	abdcgefnkh	adbceipjhg
12	M. bovis	429	1247	1221	391	282	2.2	2.2	3.2	abdcgelfkm	adbceifhjq
13	M. leprae	422	1271	1244	402	282	2.2	2.2	3.2	abcdgefkml	adbceifjhq
14	M. tuberculosis	587	1862	1823	589	358	2.0	2.2	3.3	adbcghemjk	adbjhmceit
15	B. subtilis	785	2794	2741	916	516	2.2	2.1	3.3	abdcjhmegf	adhbjcimef
16	E. faecalis	386	1244	1218	382	281	2.1	2.2	3.1	bdacgelfik	adbciefghj
17	C. acetobutylicum	494	1624	1578	511	344	2.1	2.2	3.3	abcdgefhlk	adbceijhfo
18	M. genitalium	209	535	525	196	85	2.4	2.2	3.5	bdcgzxuyos	adbcguvwos
19	M. pneumoniae	178	470	466	154	88	2.3	2.2	3.2	bcdgxoyasl	dabcgowvsr
20	S. pneumoniae	416	1331	1298	412	288	2.1	2.2	3.2	abdcgelfno	adbceifghj
21	S. pyogenes	403	1300	1277	404	280	2.1	2.2	3.1	abdcegfoln	adbceifohg
22	C. tepidum	389	1097	1062	333	231	2.1	2.2	3.3	badcgenfki	dabceipgfq
23	R. capsulatus	670	2174	2122	711	427	2.1	2.2	3.4	abcdhgefjk	adbjhicmet
24	R. prowazekii	214	510	504	155	100	2.3	2.3	3.4	bdacfegilm	dabicfemgt
25	N. gonorrhoeae	406	1298	1270	413	285	2.1	2.2	3.2	abdcgefkjh	adbiechfjg
26	N. meningitidis	381	1212	1181	380	271	2.2	2.2	3.2	abdcegfkli	adbecifhjg
27	C. jejuni	380	1142	1115	359	254	2.1	2.3	3.2	abdcegfkih	adbeicfgjh
28	H. pylori	375	1181	1144	375	246	2.0	2.3	3.3	abcdgefnhk	dabciejfhp
29	E. coli	778	2904	2859	968	570	2.2	2.1	3.2	abcdhjemlf	adhjbciefm
30	S. typhi	819	3008	2951	1007	577	2.2	2.2	3.2	abcdhjegfm	adhjbciefm
31	Y. pestis	568	1754	1715	580	386	2.1	2.2	3.3	abdcgeklfh	adbceihjfl
32	A. actinomycetemcomitans	395	1202	1166	380	271	2.1	2.2	3.2	bacdgefikl	adbciefhgj
33	H. influenzae	526	1773	1746	597	361	2.1	2.3	3.2	abcdegfjhm	adbchiefju
34	P. aeruginosa	734	2453	2398	799	490	2.1	2.2	3.3	abdchjkgef	adjhbimcef
35	T. pallidum	207	562	555	175	124	2.2	2.3	3.1	bdcgaelnfh	dabcegiplf
36	B. burgdorferi	187	442	438	140	106	2.3	2.4	3.0	bdgcaleifn	dabcgifeαl
37	T. maritima	338	1004	976	302	223	2.1	2.2	3.2	badcegfikn	dabceifgqh
38	D. radiodurans	815	2870	2811	965	557	2.2	2.1	3.3	acbdhjgekm	adhbjcimef
39	E. nidulans	383	1095	1081	339	254	2.1	2.2	3.3	abdcegjhfl	adbcjheifq
40	S. cerevisiae	561	1934	1889	596	402	2.0	2.2	3.3	abdcehjgkm	adbhcjeifm
41	C. elegans	462	1446	1418	450	295	2.1	2.2	3.3	abdcjhelgk	adbhcjiefm

42	O. sativa	292	763	751	238	178	2.1	2.3	3.5	badcegljkn	adbcehjifn
43	A. thaliana	302	804	789	250	185	2.1	2.3	3.5	badcegjhlk	adbcehjign

Summary of the characteristics of the 43 investigated organisms. For each organism we show the number of substrate (*N*), number of links (*L*), number of individual reactions or temporary substrate-enzyme complexes (*R*), number of enzymes (*E*), the exponent γ_{in} and γ_{out} and the diameter of the metabolic network (*D*). In the last two columns we list the ten substrates with the largest number of incoming (IN) and outgoing (OUT) links. The letters correspond to: a=H_2O, b=ADP, c=orthophosphate, d=ATP, e=L-glutamate, f=$NADP^+$, g=pyrophosphate, h=NAD^+, i=NADPH, j=NADH, k=CO_2, l=NH_4^+, m=CoA, n=AMP, o=pyruvate, p=L-glutamine, q=2-oxoglutarate, r='alpha'-D-glucose 1-phosphate, s=phospho`enol`pyruvate, t=acetyl-CoA, u=H^+, v=uridine, w=cytidine, x=UMP, y=CMP, z=glycerol, α=D-fructose 6-phosphate. The color code of the fields denotes the different domains of life such a magenta = Archae green = Bacterium sky blue =Eukaryote.