degradation of pollen RNA. *Nature* **347,** 757–760 (1990).
7. Matton, D. P. *et al.* Hypervariable domains of self-incompatibility RNases mediate allele-specific pollen recognition. *Plant Cell* **9,** 1757–1766 (1997).
8. Matton, D. P. *et al.* The production of an S-RNase with dual specificity suggests a novel hypothesis for the generation of new S-alleles. *Plant Cell* **11,** 2087–2097 (1999).
9. Cheung, A. The pollen tube growth pathway: its molecular and biochemical contributions and responses to pollination. *Sex. Plant Reprod.* **9,** 330–336 (1996).
10. de Nettancourt, D. *et al.* Ultrastructural aspects of the self-incompatibility mechanism in *Lycopersicum peruvianum* Mill. *J. Cell Sci.* **12,** 403–419 (1973).
11. Steer, M. W. & Steer, J. M. Pollen tube tip growth. *New Phytol.* **111,** 323–358 (1989).
12. Anderson, M. *et al.* Cloning of cDNA for a stylar glycoprotein associated with expression of self-incompatibility in *Nicotiana alata*. *Nature* **321,** 38–44 (1986).
13. Certal, A. *et al.* S-RNases in apple are expressed in the pistil along the pollen tube growth path. *Sex. Plant Reprod.* **12,** 94–98 (1999).
14. Gray, J., McClure, B., Bönig, I., Anderson, M. & Clarke, A. Action of the style product of the self-incompatibility gene of *Nicotiana alata* (S-RNase) on *in vitro* grown pollen tubes. *Plant Cell* **3,** 271–283 (1991).
15. Jahnen, W., Lush, W. & Clarke, A. Inhibition of *in vitro* pollen tube growth by isolated S-glycoproteins of *Nicotiana alata*. *Plant Cell* **1,** 501–510 (1989).
16. Hess, D., Gresshoff, P., Fielitz, U. & Gleiss, D. Uptake of protein and bacteriophage into swelling and germinating pollen of *Petunia hybrida*. *Z. Pflanzenphysiol.* **74,** 371–376 (1974).
17. Golz, J., Clarke, A. E. & Newbegin, E. Mutational approaches to the study of self-incompatibility: revisiting the pollen-part mutants. *Ann. Botany* **85,** 95–103 (2000).
18. Lewis, D. Chromosome fragments and mutation of the incompatibility gene. *Nature* **190,** 990–991 (1961).
19. Pandey, K. Elements of the S-gene complex. II. Mutations and complementation at the SI locus in *Nicotiana alata*. *Heredity* **22,** 255–284 (1967).
20. Saba-El-Leil, M., Rivard, S., Morse, D. & Cappadocia, M. The S11 and S13 self incompatibility alleles in *Solanum chacoense* Bitt. are remarkably similar. *Plant Mol. Biol.* **24,** 571–583 (1994).
21. Despres, C., Saba-El-Leil, M., Rivard, S., Morse, D. & Cappadocia, M. Molecular cloning of two *Solanum chacoense* S-alleles and a hypothesis concerning their evolution. *Sex. Plant Reprod.* **7,** 169–176 (1994).
22. Van Sint Jan, V., Laublin, G., Birhman, R. & Cappadocia, M. Genetic analysis of leaf explant regenerability in *Solanum chacoense*. *Plant Cell Tissue Org. Cult.* **47,** 9–13 (1996).
23. Veronneau, H., Lavoie, G. & Cappadocia, M. Genetic analysis of anther and leaf disk culture in two clones of *Solanum chacoense* Bitt and their reciprocal hybrids. *Plant Cell Tissue Org. Cult.* **30,** 199–209 (1992).
24. Rivard, S., Saba-El-Leil, M., Landry, B. & Cappadocia, M. RFLP analyses and segregation of molecular markers in plants produced by *in vitro* anther culture, selfing, and reciprocal crosses of two lines of self-incompatible *Solanum chacoense*. *Genome* **37,** 775–783 (1994).
25. Vandenbosch, K. in *Electron Microscopy of Plant Cells* (eds Hall, J. & Hawes, C.) 181–218 (Academic, New York, 1991).
26. Martin, F. Staining and observing pollen tubes in the style by means of fluorescence. *Stain Technol.* **34,** 125–128 (1959).
27. McFadden, G. in *Electron Microscopy of Plant Cells* (eds Hall, J. & Hawes, C.) 220–255 (Academic, New York, 1991).

**Acknowledgements**

We thank L. Pelletier and N. Nassoury for technical assistance, G. Teodorescu for plant care, and S. McCormick, A. Cheung, T.-H. Kao and V. de Luca for helpful discussions. The work was supported by a fellowship from Programme Québecois de Bourses d'Excellence, Québec (D.-T.L.) and by grants from NSERC (M.C.) and Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (D.M., M.C.).

Correspondence and requests for materials should be addressed to M.C. (e-mail: mario.cappadocia@umontreal.ca).

# The large-scale organization of metabolic networks

**H. Jeong*, B. Tombor†, R. Albert*, Z. N. Oltvai† & A.-L. Barabási***

* *Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA*
† *Department of Pathology, Northwestern University Medical School, Chicago, Illinois 60611, USA*

**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 cellular constituents and reactions[1]. However, despite the key role of these networks in sustaining cellular functions, their large-scale structure is essentially unknown. Here we present a systematic comparative mathematical analysis of the metabolic networks of 43 organisms representing all three domains of life. We show that, despite significant variation in their individual constituents and pathways, these metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex non-biological systems[2]. This may indicate that metabolic organization is not only identical for all living organisms, but also complies with the design principles of robust and error-tolerant scale-free 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 is rooted in the dynamic interactions among its many constituents[14–16], such as proteins, DNA, RNA and small molecules. Scientific developments have improved our ability to identify the design principles that integrate these interactions into a complex system. Large-scale sequencing projects have not only provided complete sequence information for a number of genomes, but also allowed the development of integrated pathway–genome databases[17–19] that provide organism-specific connectivity maps of metabolic and, to a lesser extent, other cellular networks. However, owing to the large number and diversity of the constituents and reactions that form such networks, these maps are extremely complex, offering only limited insight into the organizational principles of these systems. Our ability to address in quantitative terms the structure of these cellular networks has benefited from advances in understanding the generic properties of complex networks[2].

Until recently, complex networks have been modelled using the classical random network theory introduced by Erdös and Rényi[20,21]. The Erdös–Rényi model assumes that each pair of nodes (that is, 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 that peaks strongly at $\langle k \rangle$ (Fig. 1b), implying that the probability of finding a highly connected node decays exponentially ($P(k) \approx e^{-k}$ for $k \gg \langle k \rangle$). On the other hand, 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, showing that these systems are described by scale-free networks[2] (Fig. 1c), for which $P(k)$ follows a power-law, $P(k) \approx k^{-\gamma}$ (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). As the distinction between scale-free and exponential networks emerges as a result of simple dynamical principles[24,25], understanding the large-scale structure of cellular networks can not only provide 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 begin to address the large-scale structural organization of cellular networks, we have examined the topological properties of the core metabolic network of 43 different organisms based on data deposited in the WIT database[19]. This integrated pathway–genome database predicts the existence of a given metabolic pathway on the basis of the annotated genome of an organism combined with firmly established data from the biochemical literature. As 18 of the 43 genomes deposited in the database are not yet fully sequenced, and a substantial portion of the identified open reading frames are

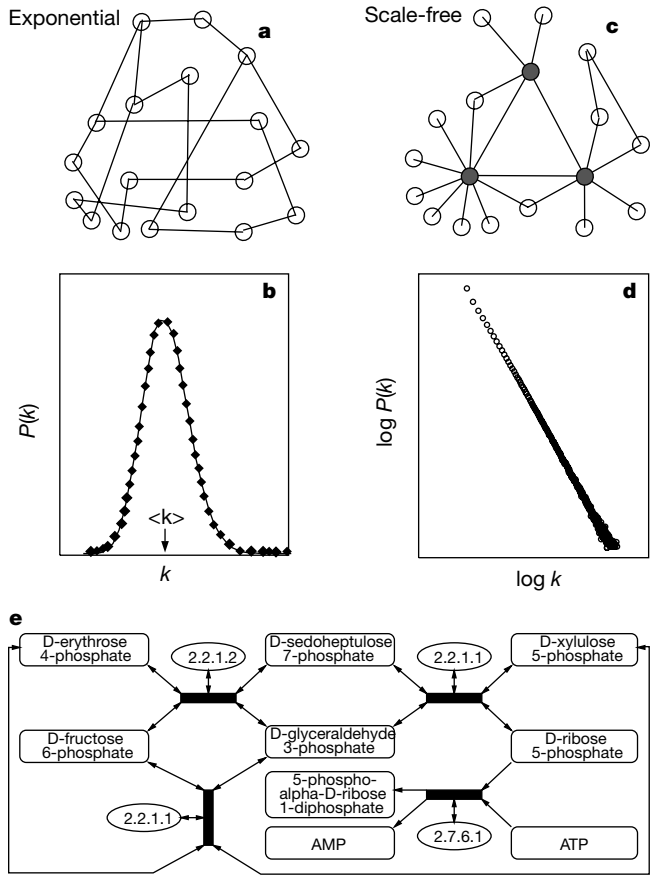



**Figure 1** Attributes of generic network structures. **a**, Representative structure of the network generated by the Erdös–Rényi network model[20,21]. **b**, The network connectivity can be characterized by the probability, $P(k)$, that a node has $k$ links. For a random network $P(k)$ peaks strongly at $k = \langle k \rangle$ and decays exponentially for large $k$ (that is, $P(k) \approx e^{-k}$ 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 (red), 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, $P(k) \approx k^{-\gamma}$, appearing as a straight line with slope $-\gamma$ on a log–log plot. **e**, A portion of the WIT database for *E. coli*. Each substrate can be represented as a node of the graph, linked through temporary educt–educt complexes (black boxes) from which the products emerge as new nodes (substrates). The enzymes, which provide the catalytic scaffolds for the reactions, are shown by their EC numbers.

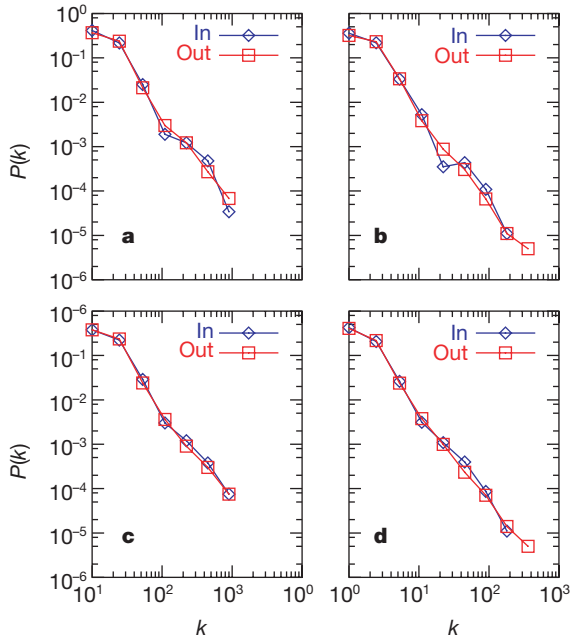



**Figure 2** Connectivity distributions $P(k)$ for substrates. **a**, *Archaeoglobus fulgidus* (archae); **b**, *E. coli* (bacterium); **c**, *Caenorhabditis elegans* (eukaryote), shown on a log–log plot, counting separately the incoming (In) and outgoing links (Out) for each substrate. $k_{in}$ ($k_{out}$) corresponds to the number of reactions in which a substrate participates as a product (educt). The characteristics of the three organisms shown in **a**–**c** and the exponents $\gamma_{in}$ and $\gamma_{out}$ for all organisms are given in Table 1 of the Supplementary Information. **d**, The connectivity distribution averaged over all 43 organisms.

functionally unassigned, the list of enzymes, and consequently the list of substrates and reactions (see Table 1 in Supplementary Information), will certainly be expanded in the future. Nevertheless, this publicly available database represents our best approximation for the metabolic pathways in 43 organisms and provides sufficient data for their unambiguous statistical analysis (see Methods and Supplementary Information).

As we show in Fig. 1e, we 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, 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. This representation allows us systematically to 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: that is, to establish whether their topology is best described by the inherently random and uniform exponential model[21] (Fig. 1a, b), or the highly heterogeneous scale-free model[2] (Fig. 1c, 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; in other words, metabolic networks belong to the class of scale-free networks. As under physiological conditions a large number of biochemical reactions (links) in a metabolic network are preferentially catalysed 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 $k$ metabolic reactions follows $P(k) \approx k^{-\gamma_{in}}$, 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; see Supplementary Information, also available at www.nd.edu/~networks/cell), indicating the generic nature of this structural organization (Fig. 2d).

A general feature of many complex networks is their small-world character[26], meaning that 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, the average connectivity of a node is fixed, which implies that the diameter of a network increases logarithmically with the addition of new nodes[2,26,27]. For metabolic networks this implies that a more complex bacterium with more enzymes and substrates, such as *E. coli*, would have a larger diameter than a simple bacterium, such as *Mycoplasma genitalium*. We find, however, that the diameter of the metabolic network is the same for all 43 organisms,

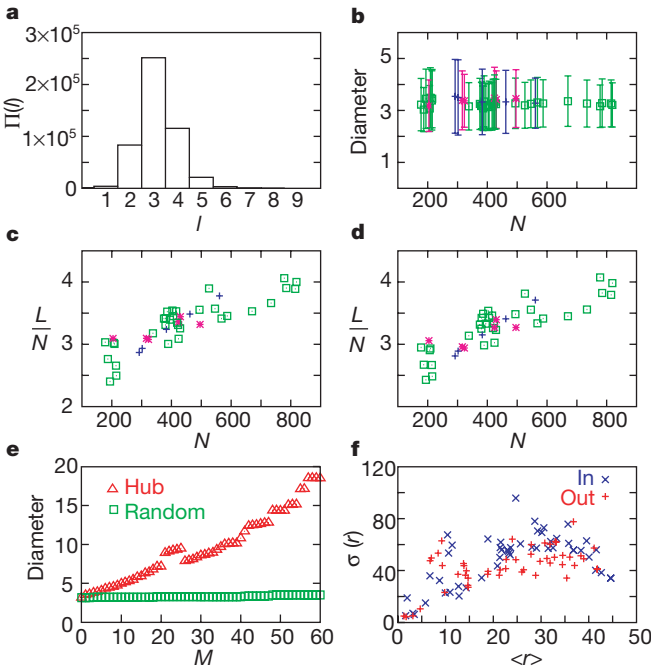



**Figure 3** Properties of metabolic networks. **a**, The histogram of the biochemical pathway lengths, $l$, in *E. coli*. **b**, The average path length (diameter) for each of the 43 organisms. Error bars represent standard deviation $\sigma \approx \langle l^2 \rangle - \langle l \rangle^2$ as determined from $\Pi(l)$ (shown in **a** for *E. coli*). **c**, **d**, Average number of incoming links (**c**) or outgoing links (**d**) per node for each organism. **e**, The effect of substrate removal on the metabolic network diameter of *E. coli*. In the top curve (red) the most connected substrates are removed first. In the bottom curve (green) nodes are removed randomly. $M = 60$ corresponds to ~8% of the total number of substrates in found in *E. coli*. **f**, Standard deviation of the substrate ranking ($\sigma_r$) as a function of the average ranking $\langle r \rangle_0$ for substrates present in all 43 organisms investigated. The horizontal axis in **b**–**d** denotes the number of nodes in each organism. **b**–**d**, Archaea (magenta), bacteria (green) and eukaryotes (blue) are shown.

irrespective of the number of substrates found in the given species (Fig. 3b). This is unexpected, and is possible only if with increasing organism complexity individual substrates are increasingly connected to maintain a relatively constant metabolic network diameter. We find that the average number of reactions in which a certain substrate participates increases with the number of substrates found within a given organism (Fig. 3c, 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. But scale-free networks also demonstrate unexpected robustness against random errors[5]. To investigate whether metabolic networks display a similar error tolerance we performed computer simulations on the metabolic network of *E. coli*. Upon removal of the most connected substrates the diameter increases rapidly, illustrating the special role of these metabolites in maintaining a constant metabolic network diameter (Fig. 3e). However, when a randomly chosen $M$ substrates are removed—mimicking the consequence of random mutations of catalysing enzymes—the average distance between the remaining nodes is not affected, indicating a striking insensitivity to random errors. Indeed, *in silico* and *in vivo* mutagenesis studies indicate remarkable fault tolerance upon removal of a substantial number of metabolic enzymes from the *E. coli* metabolic network[28]. Data similar to those shown in Fig. 3e have been obtained for all organisms investigated, without detectable correlations with their evolutionary position.

As the large-scale architecture of the metabolic network rests on the most highly connected substrates, we need to investigate whether the same substrates act as hubs in all organisms, or whether there are organism-specific differences in the identity of the most connected substrates. When we rank all the substrates in a given organism on the basis of the number of links they have (Table 1; see Supplementary Information), we find that the ranking of the most connected substrates is practically identical for all 43 organisms. Also, only around 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 organisms emerge for less connected substrates. To quantify this observation, we examined the standard deviation ($\sigma_r$) of the rank for substrates that are present in all 43 organisms. As shown in Fig. 3f, $\sigma_r$ 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 the educts or products of species-specific enzymatic activities.

The contemporary topology of a metabolic network reflects a long evolutionary process moulded in general for a robust response towards internal defects and environmental fluctuations and in particular to the ecological niche occupied by a specific organism. As a result, we would expect that these networks are far from random, and our data show that the large-scale structural organization of metabolic networks is indeed very similar to that of robust and error-tolerant networks[2,5]. The uniform network topology observed in all 43 organisms indicates that, irrespective of their individual building blocks or species-specific reaction pathways, the large-scale structure of metabolic networks may be identical in all living organisms, in which the same highly connected substrates may provide the connections between modules responsible for distinct metabolic functions[1].

A unique feature of metabolic networks, as opposed to 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, as a larger diameter would attenuate the organism's ability to respond efficiently to external changes or internal errors. For example, if the concentration of a substrate were to suddenly diminish owing to a mutation in its main catalysing enzyme, offsetting the changes would involve the activation of longer alternative biochemical pathways, and consequently the synthesis of more new enzymes, than within a metabolic network with a smaller diameter.

How generic are these principles for other cellular networks (for example, apoptosis or cell cycle)? Although the current mathematical tools do not allow unambiguous statistical analysis of the topology of other networks owing to their relatively small size, our preliminary analysis indicates that connectivity distribution of non-metabolic pathways may also follow 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 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[19] (http://igweb.integratedgenomics.com/IGwit/), which predicts the existence of a metabolic pathway in an organism on the basis of its annotated genome (on the presence of the presumed open reading frame of an enzyme that catalyses a given metabolic reaction), in combination with firmly established data

from the biochemical literature. As of December 1999, this database provides descriptions for 6 archaea, 32 bacteria and 5 eukaryotes. 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. Bidirectional reactions were considered separately. For a given organism with $N$ substrates, $E$ enzymes and $R$ intermediate complexes the full stoichiometric 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, showing how many substrates have exactly $k_{in} = 0,1,\ldots$. 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, $\Pi(l)$ (that is, the smallest number of reactions by which one can reach substrate B from substrate A) was determined using a burning algorithm. From $\Pi(l)$ we determined the diameter, $D = \Sigma_l l\cdot\Pi(l)/\Sigma_l\Pi(l)$, which represents the average path length between any two substrates.

### Substrate ranking $\langle r\rangle_o$, $\sigma(r)$

Substrates present in all 43 organisms (a total of 51 substrates) were ranked on the basis of the number of links each had in each organisms, having considered incoming and outgoing links separately ($r = 1$ was assigned for the substrate with the largest number of connections, $r = 2$ for the second most connected one, and so on). This gave a well defined $r$ value in each organism for each substrate. 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 analysed, the genomes of 25 have been completely sequenced (5 archaea, 18 bacteria and 2 eukaryotes), whereas the remaining 18 are only partially sequenced. Therefore two main sources of possible errors in the database could affect our analysis: the erroneous annotation of enzymes and, consequently, biochemical reactions (the likely source of error for the organisms with completely sequenced genomes); and reactions and pathways missing from the database (for organisms with incompletely sequenced genomes, both sources of error are possible). We investigated the effect of database errors on the validity of our findings. The data, presented in Supplementary Information, indicate that our results are robust to these errors.

Received 3 April; accepted 18 July 2000.

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,** 509–512 (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,** 1677–1679 (1999).
4. Banavar, J. R., Maritan, A. & Rinaldo, A. Size and form in efficient transportation networks. *Nature* **399,** 130–132 (1999).
5. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406,** 378–382 (2000).
6. Ingber, D. E. Cellular tensegrity: defining new rules of biological design that govern the cytoskeleton. *J. Cell Sci.* **104,** 613–627 (1993).
7. Bray, D. Protein molecules as computational elements in living cells. *Nature* **376,** 307–312 (1995).
8. McAdams, H. H. & Arkin, A. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15,** 65–69 (1999).
9. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403,** 339–342 (2000).
10. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403,** 335–338 (2000).
11. Hasty, J., Pradines, J., Dolnik, M. & Collins, J. J. Noise-based switches and amplifiers for gene expression. *Proc. Natl Acad. Sci. USA* **97,** 2075–2080 (2000).
12. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405,** 590–593 (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,** 913–917 (1997).
15. Yi, T. M., Huang, Y., Simon, M. I. & Doyle, J. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc. Natl Acad. Sci. USA* **97,** 4649–4653 (2000).
16. Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science* **283,** 381–387 (1999).
17. Karp, P. D., Krummenacker, M., Paley, S. & Wagg, J. Integrated pathway–genome databases and their role in drug discovery. *Trends Biotechnol.* **17,** 275–281 (1999).
18. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
19. Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28,** 123–125 (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, 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. *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. (cited 31 January 2000) ⟨http://xxx.lanl.gov/abs/cond-mat/0001458⟩ (2000).
25. Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of reference networks with aging (cited 28 January 2000) ⟨http://xxx.lanl.gov/abs/cond-mat/0001419⟩ (2000).
26. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393,** 440–442 (1998).
27. Barthelemy, M. & Amaral, L. A. N. Small-world networks: Evidence for a crossover picture. *Phys. Rev. Lett.* **82,** 3180–3183 (1999).
28. Edwards, J. S. & Palsson, B. O. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA* **97,** 5528–5533 (2000).

**Supplementary information** is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

### Acknowledgements

We thank all members of the WIT project for making this invaluable database publicly available. 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. (e-mail: alb@nd.edu) or Z.N.O. (e-mail: zno008@northwestern.edu).

## errata

# Determining multiple length scales in rocks

**Yi-Qiao Song, Seungoh Ryu & Pabitra N. Sen**

*Nature* **406,** 178–181 (2000).

On page 179 of this paper, the six occurrences of $\pi 2$ on lines 10 and 23 of the text should have been $\pi/2$. □

# Glycosyltransferase activity of Fringe modulates Notch–Delta interactions

**Katja Brücker, Lidia Perez, Henrik Clausen & Stephen Cohen**

*Nature* **406,** 411–415 (2000).

In Fig. 1b, the fifth column of the Fng–myc row should have shown a plus sign instead of a minus sign. □

# Supplementary material I

## Table 1.

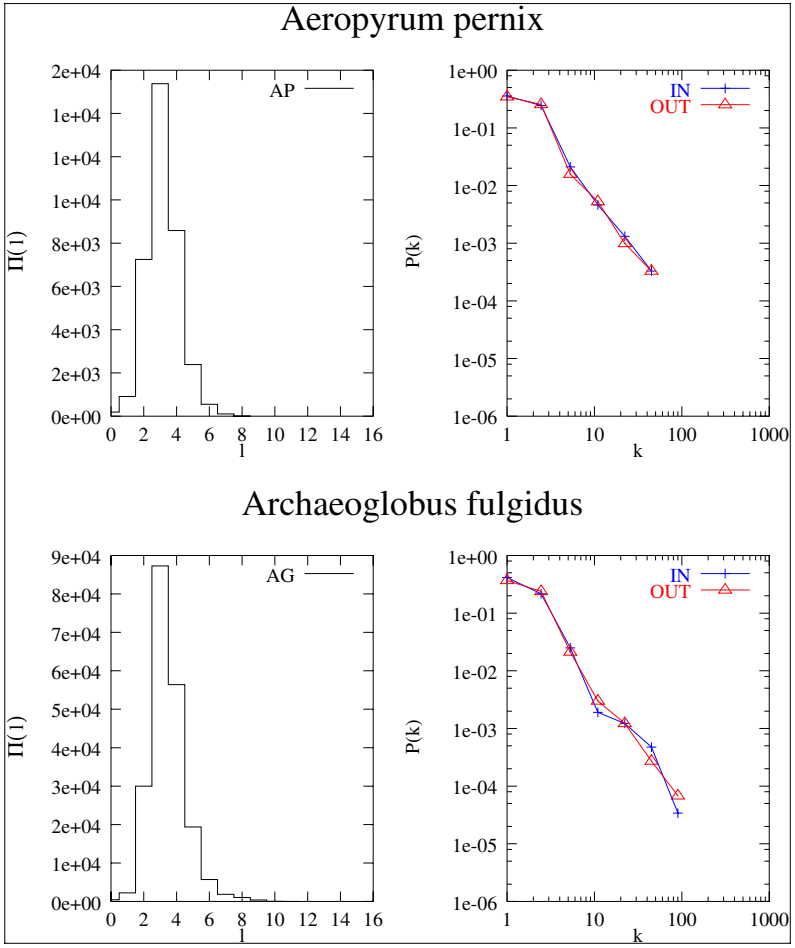
Summary of the characteristics of the 43 investigated organisms. For each organism we show the number of substrate (*N*), number of links (*L*), number of individual reactions or temporary substrate-enzyme complexes (*R*), number of enzymes (*E*), the exponent $\gamma_{in}$ and $\gamma_{out}$ and the 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.

| No. | Name | *N* | *L(IN)* | *L(OUT)* | *R* | *E* | $\gamma_{in}$ | $\gamma_{out}$ | *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. thermoautotro phicum | 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. actinomycetem comitans | 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 |

# Supplementary material II

The attached graphs show the path length distribution Π(*l*) (see Fig. 3a in the manuscript) and the connectivity distribution *P(k)* (see Fig. 2 in the manuscript) for each of the 43 investigated organisms.

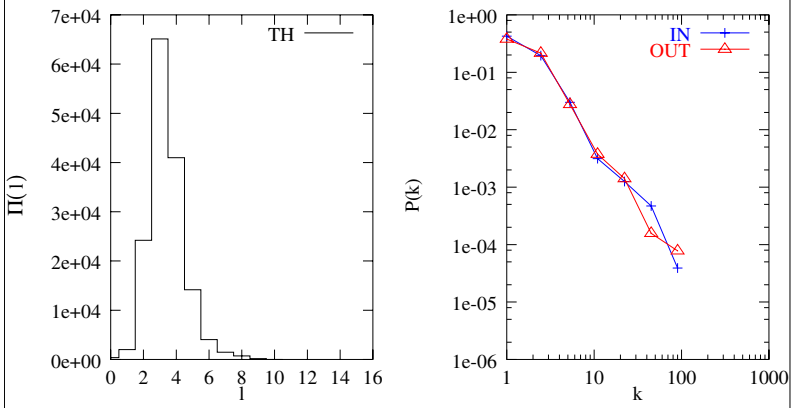
## Aeropyrum pernix



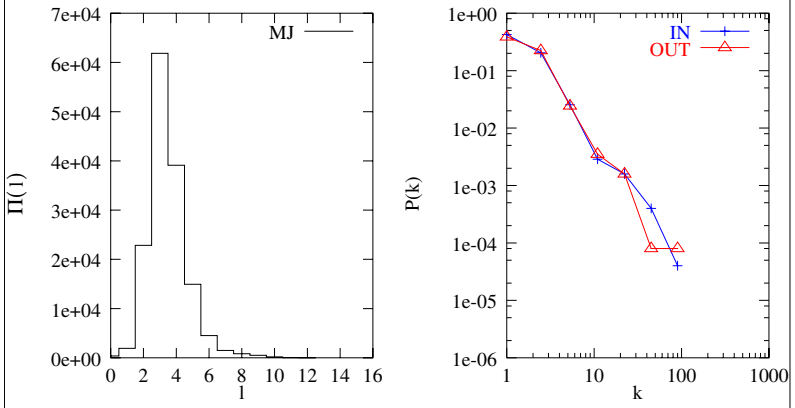

## Archaeoglobus fulgidus



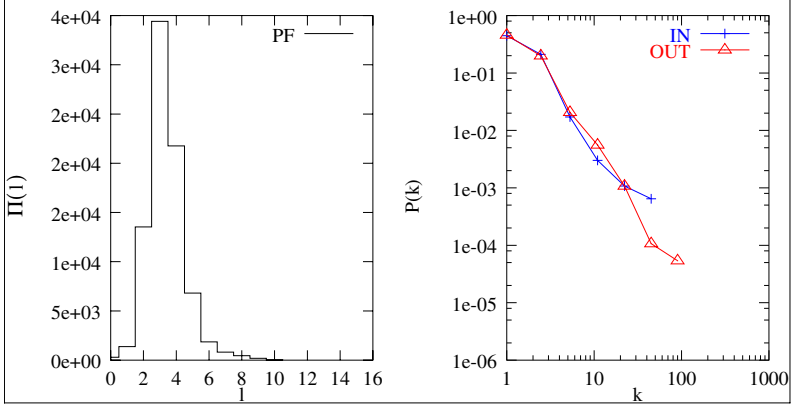

## Methanobacterium thermoautotrophicum



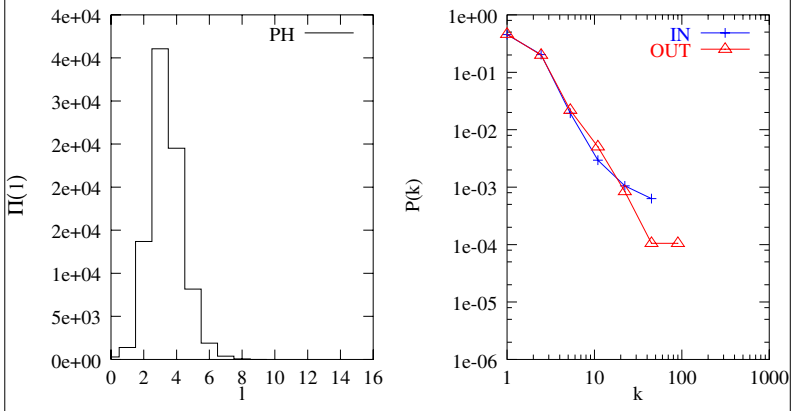

## Methanococcus jannaschii



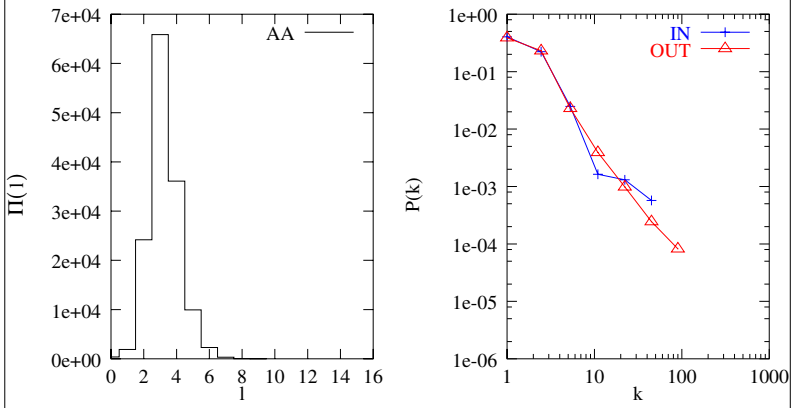

## Pyrococcus furiosus



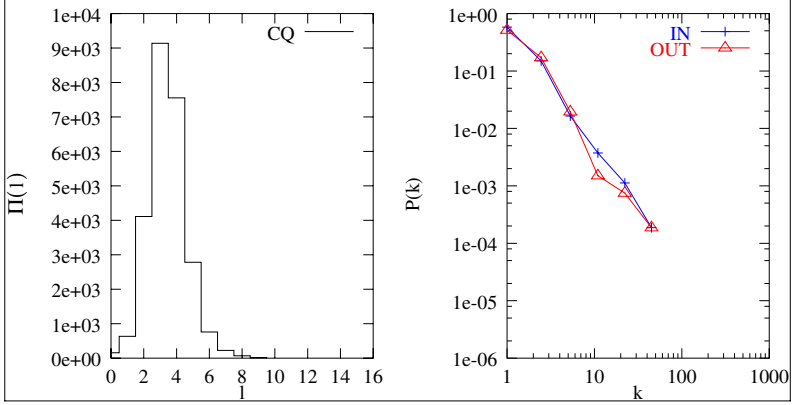
## Pyrococcus horikoshii



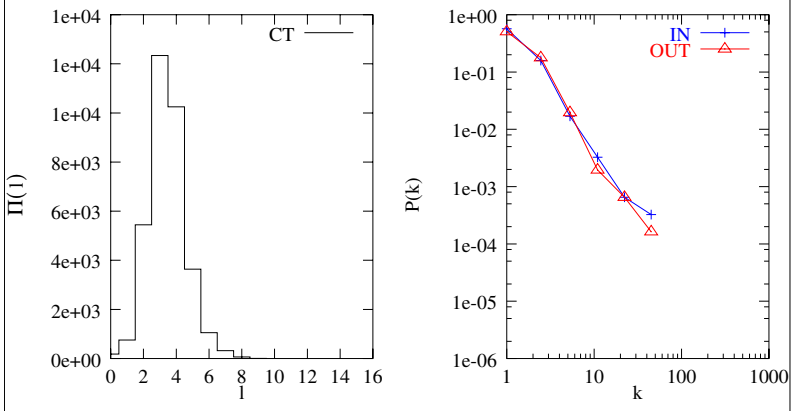

## Aquifex aeolicus



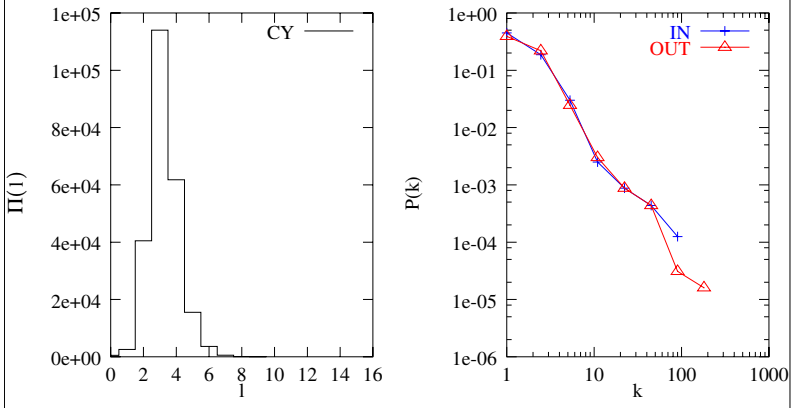

## Chlamydia pneumoniae



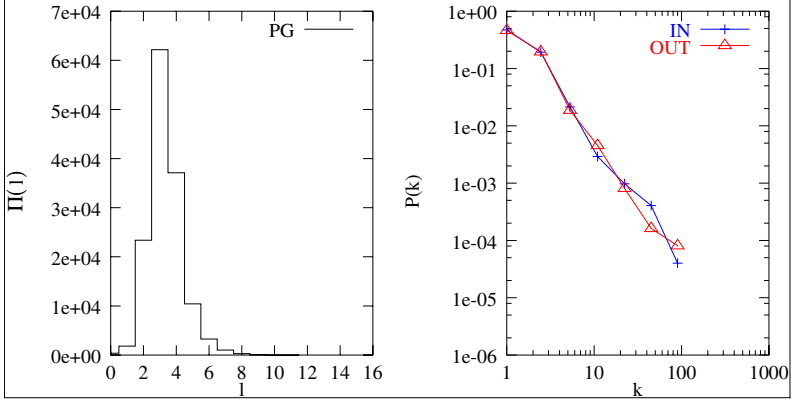

## Chlamydia trachomatis



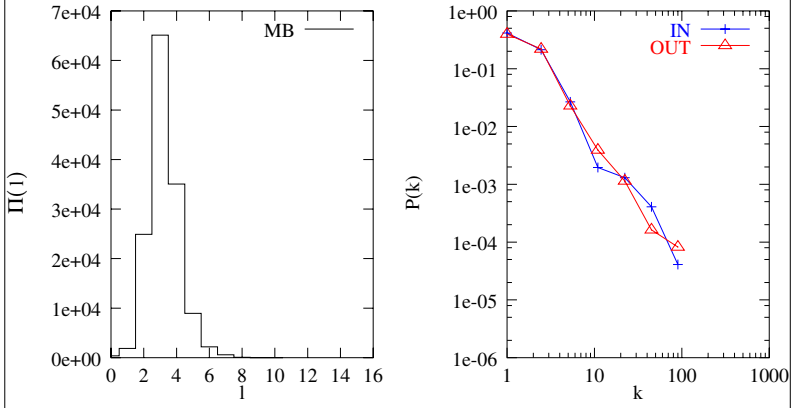

## Synechocystis sp.



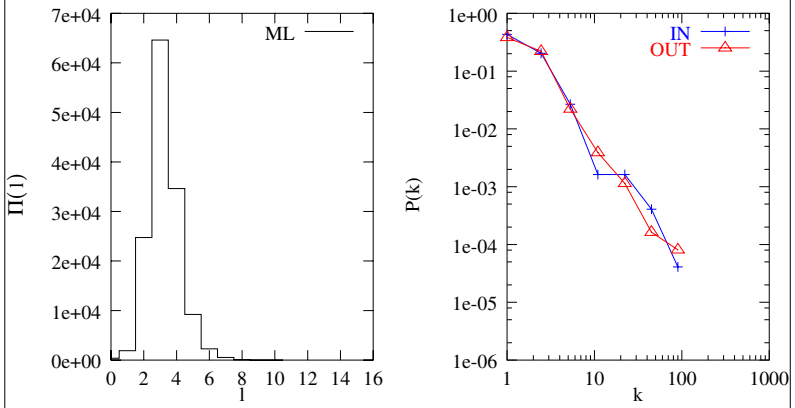

## Porphyromonas gingivalis



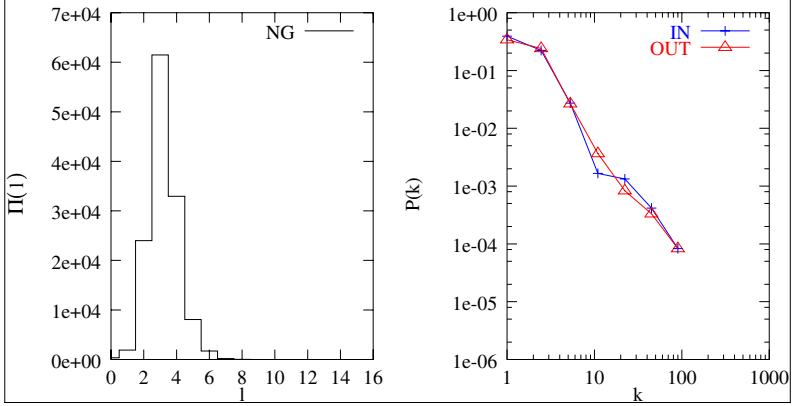
## Mycobacterium bovis



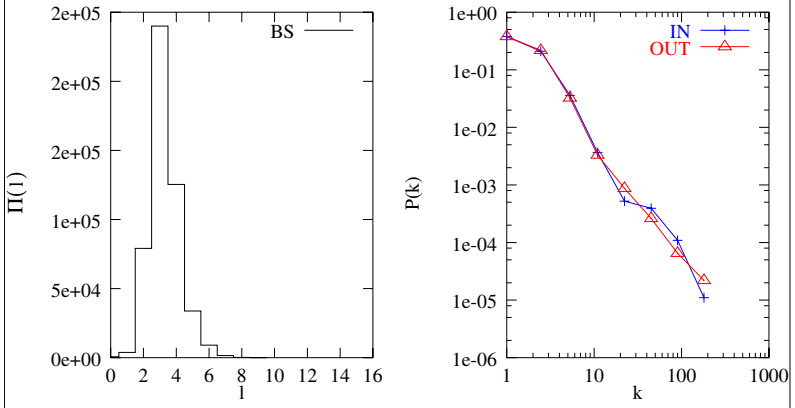

## Mycobacterium leprae



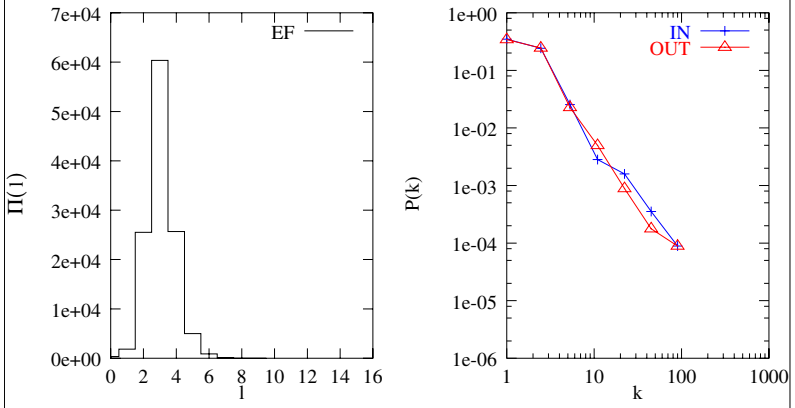

## Neisseria gonorrhoeae



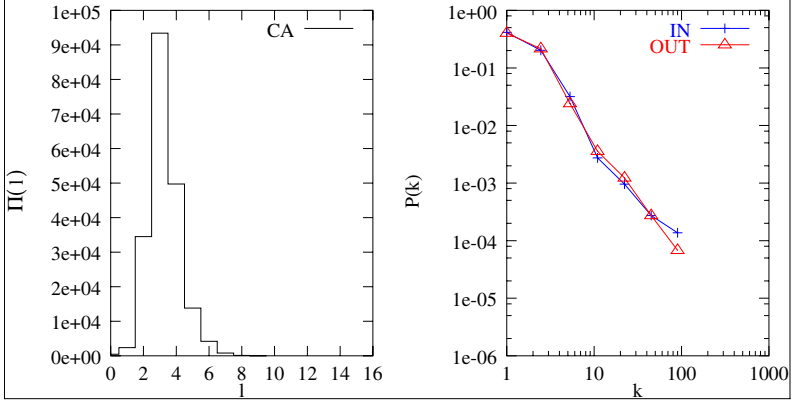

## Bacillus subtilis



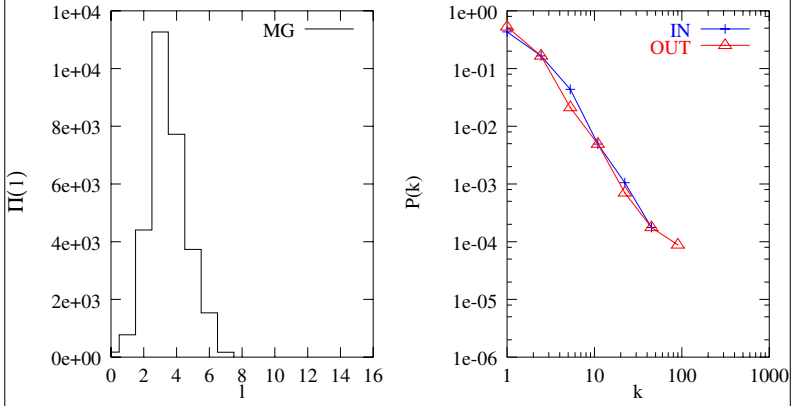

## Enterococcus faecalis



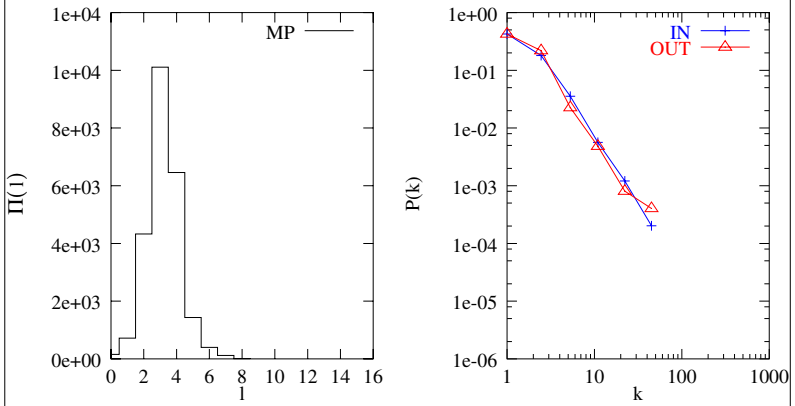

## Clostridium acetobutylicum



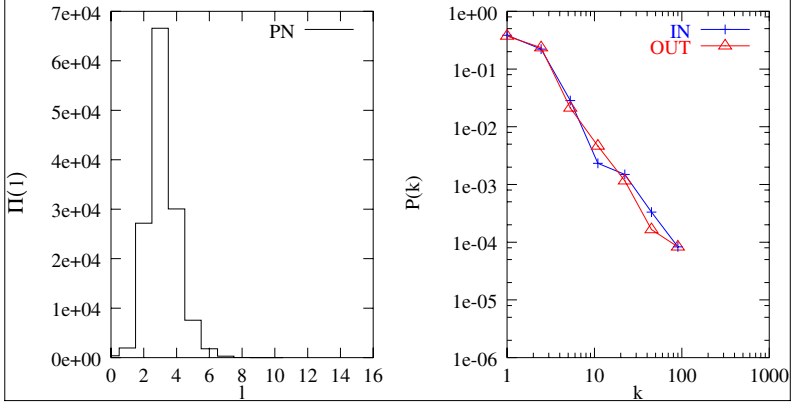
## Mycoplasma genitalium



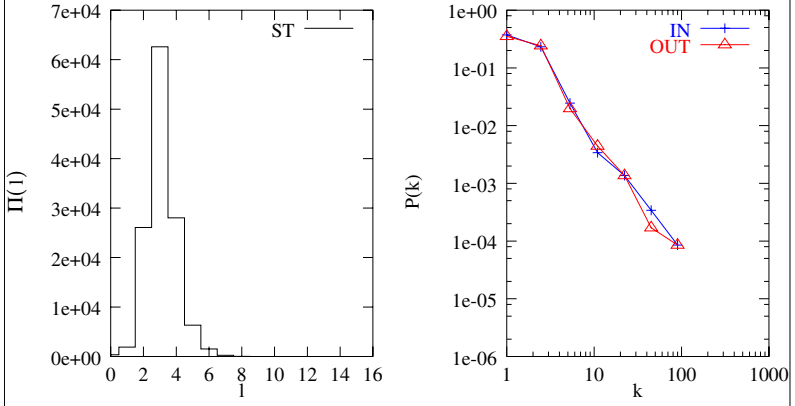

## Mycoplasma pneumoniae



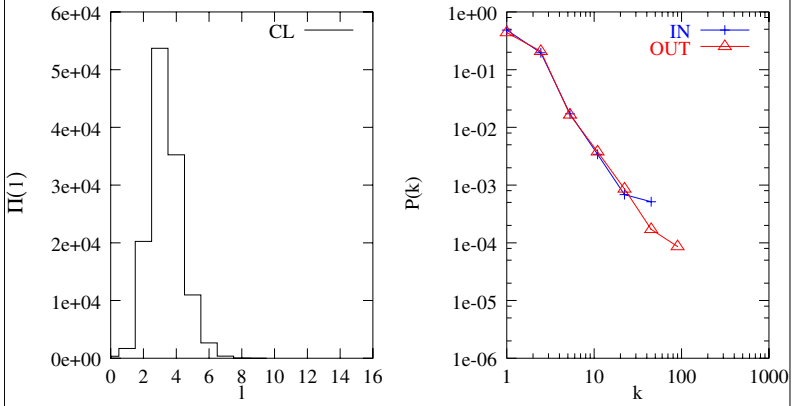

## Streptococcus pneumoniae



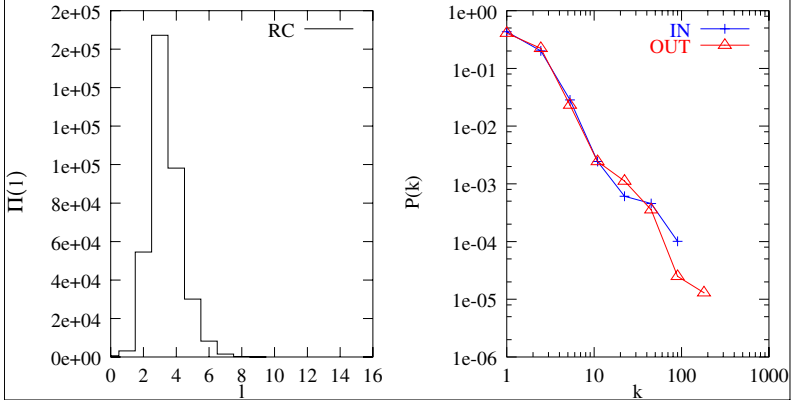

## Streptococcus pyogenes



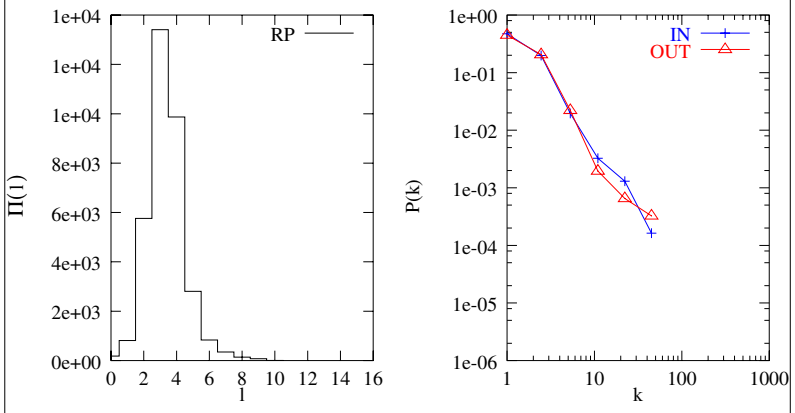

## Chlorobium tepidum



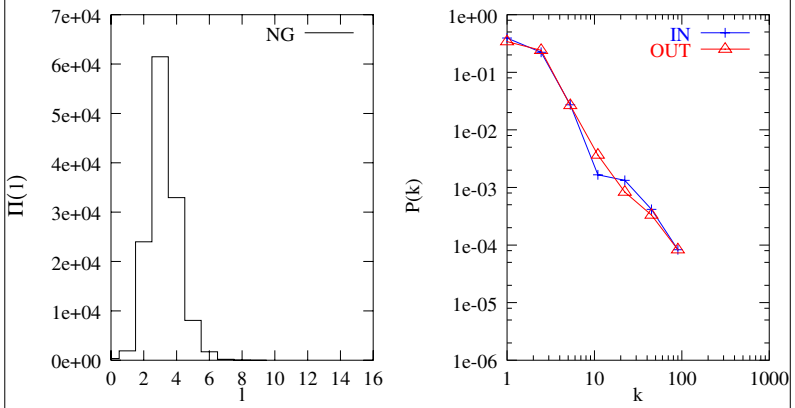

## Rhodobacter capsulatus



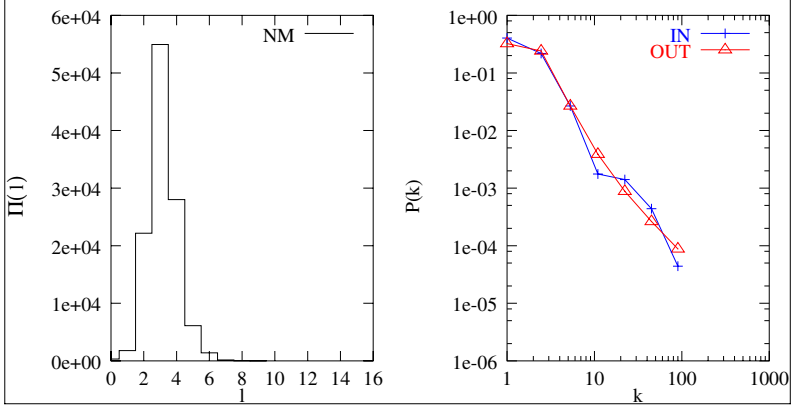
## Rickettsia prowazekii



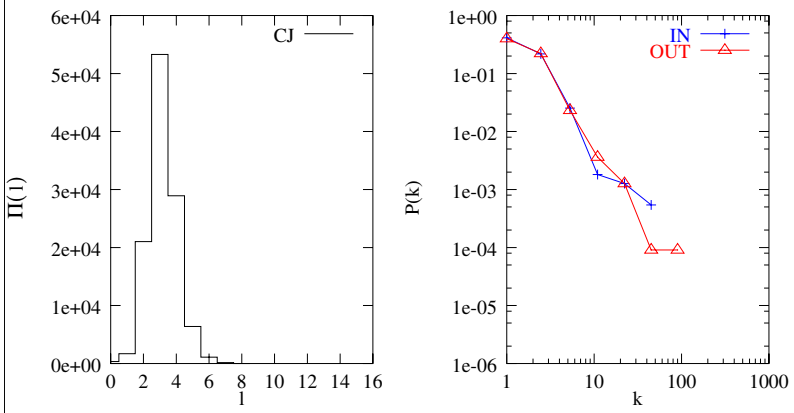

## Neisseria gonorrhoeae



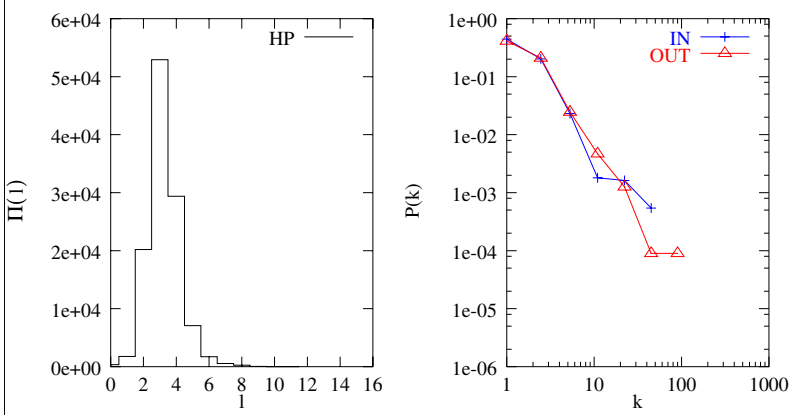

## Neisseria meningitidis



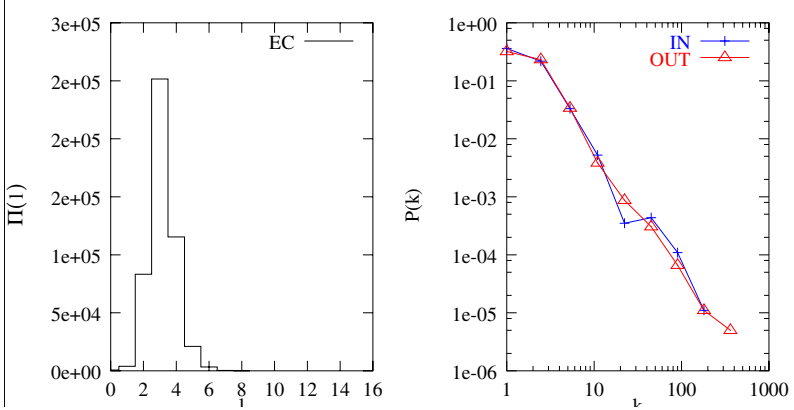

## Campylobacter jejuni



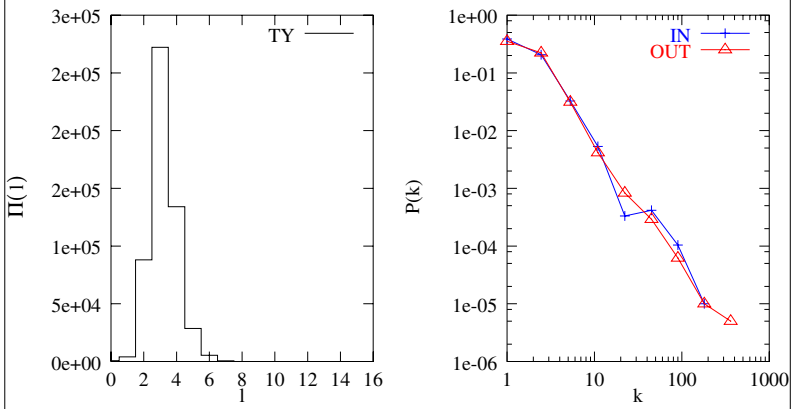

## Helicobacter pylori



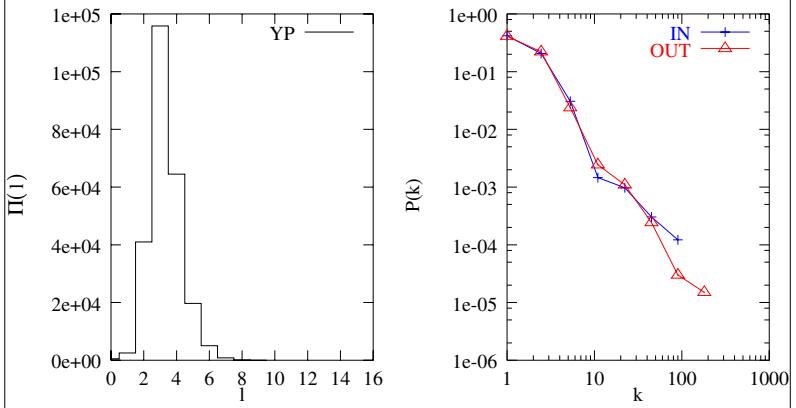

## Escherichia coli



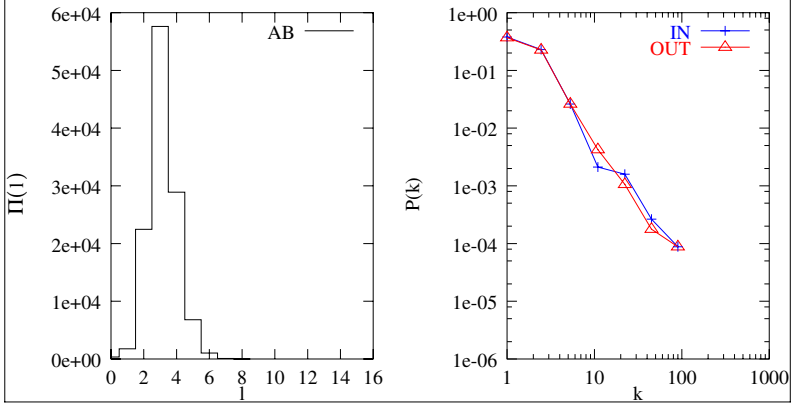
## Salmonella typhi



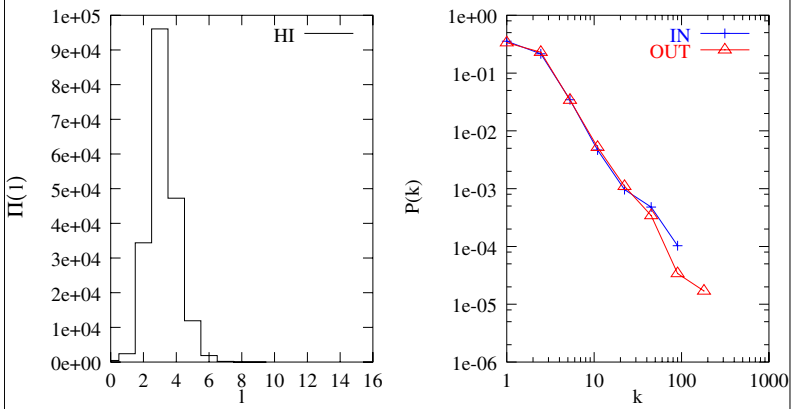

## Yersinia pestis



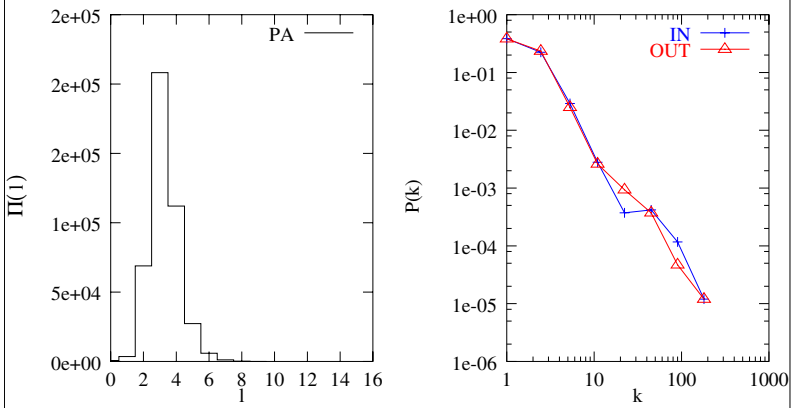

## Actinobacillus actinomycetemcomitans



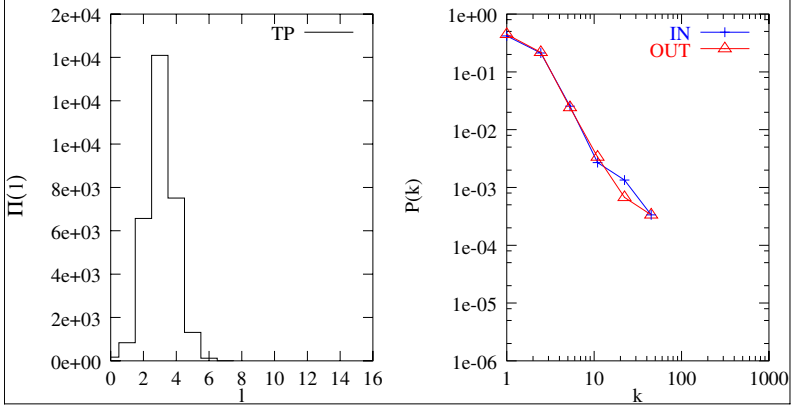

## Haemophilus influenzae



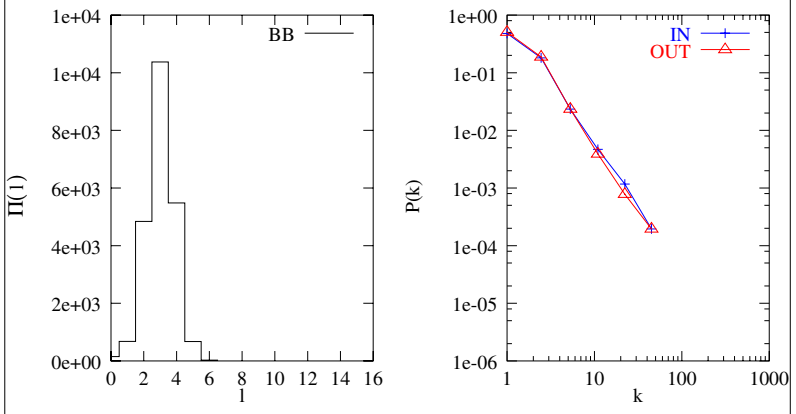

## Pseudomonas aeruginosa



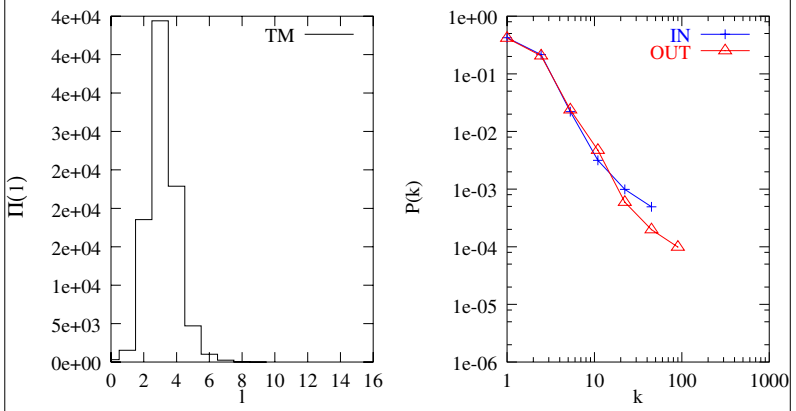

## Treponema pallidum



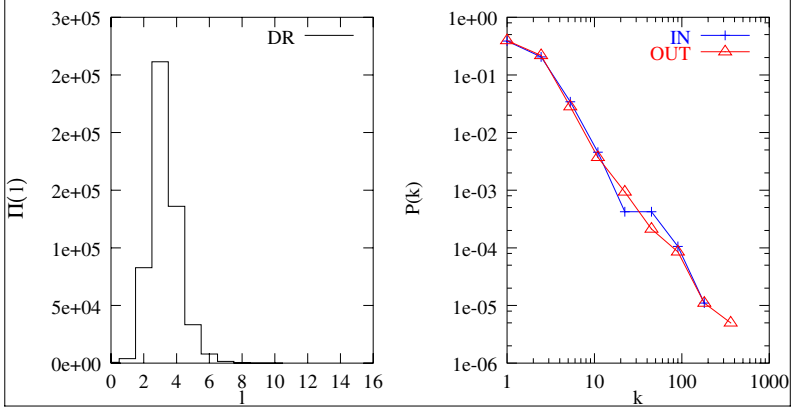
## Borrelia burgdorferi



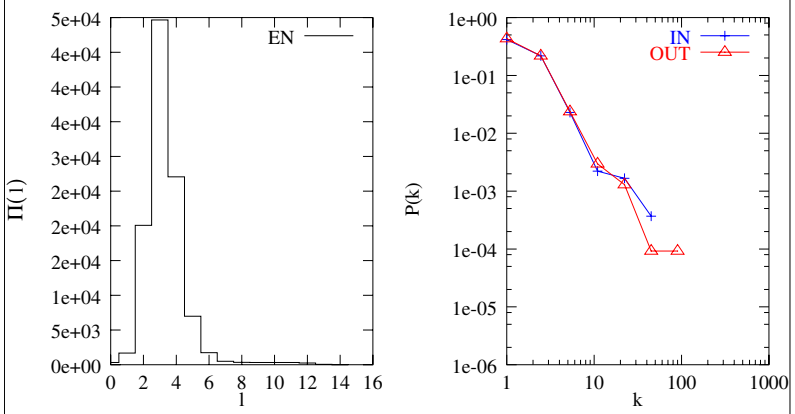

## Thermotoga maritima



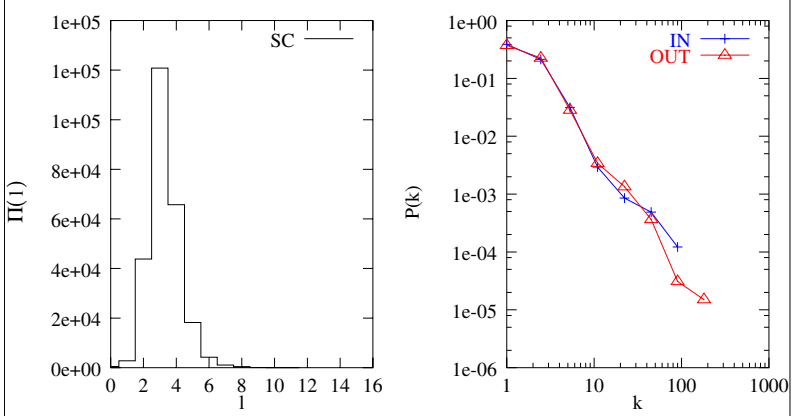

## Deinococcus radiodurans



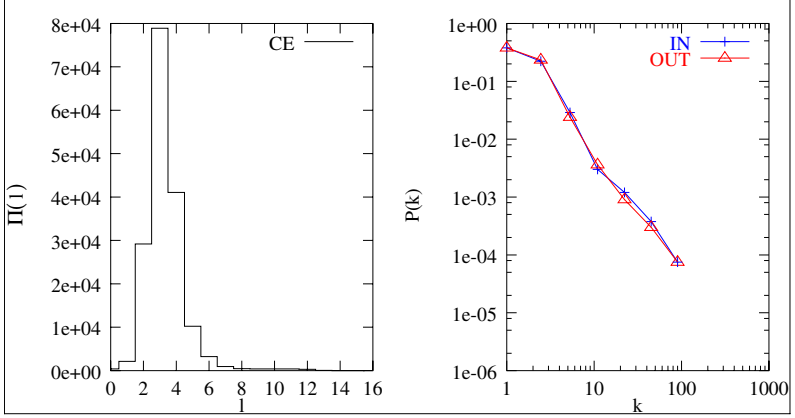

## Emericella nidulans

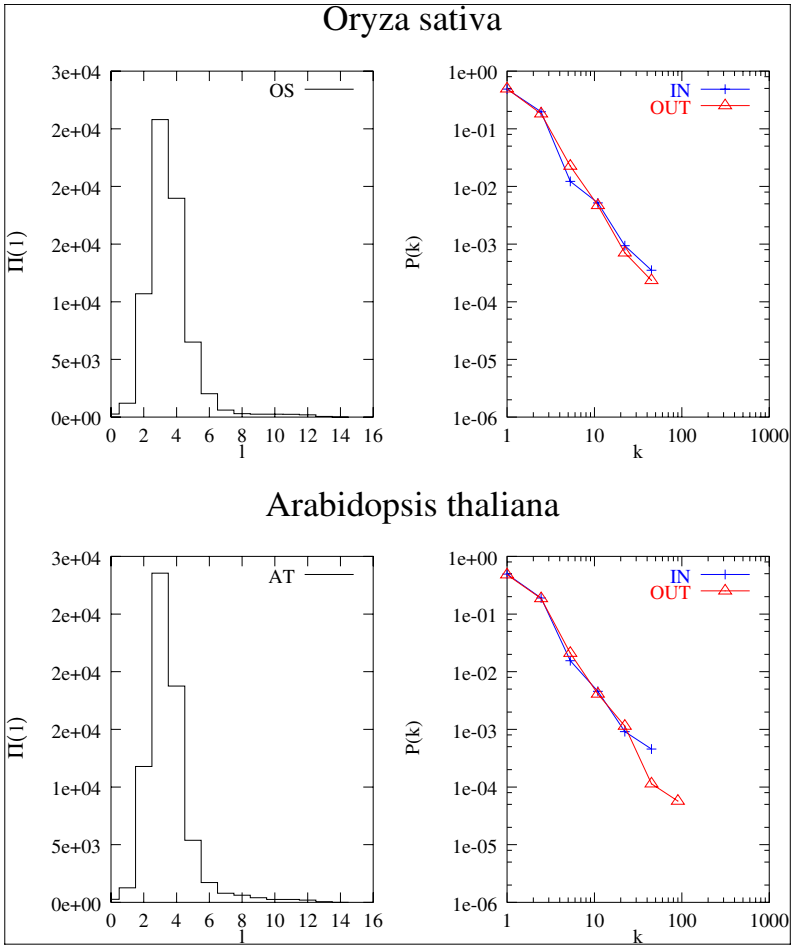



## Saccharomyces cerevisiae




## Caenorhabditis elegans




Oryza sativa
OS
Π(1)
l
P(k)
k
IN
OUT
Arabidopsis thaliana
AT
Π(1)
l
P(k)
k
IN
OUT

# Supplementary material III

## Methods

### I. Database construction.

**For our analysis the WIT database (http://igweb.integratedgenomics.com/IGwit/) was utilized. This database divides the full cellular network of each organism into 6 subgroups:**

1. Intermediate metabolism and bioenergetics
2. Information pathway
3. Electron transport
4. Transmembrane transport
5. Signal transduction
6. Structure and function of cell

For our analyses of core cellular metabolisms the Intermediate metabolism and bioenergetics portion (subgroup #1) of the WIT database was used. As of December 1999, this comprehensive publicly available integrated pathway-genome database provides description for 6 archaea, 32 bacteria and 5 eukaryota, of which 5 of 6, 18 of 32, and 2 of 5 are fully sequenced, respectively.

In the attached figure we show a typical example of a pathway.
(fructose_6-phosphate,_glyceraldehyde_3-phosphate--5-phosphoribose_1-diphosphate_anabolism)

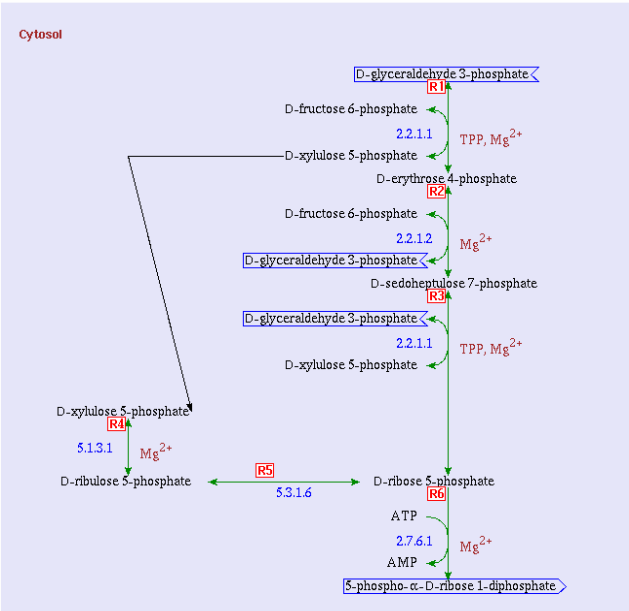



Figure M1.

Corrections applied to the WIT database:

Prior to our analysis, the downloaded data was carefully examined for inconsistencies by the following steps.

1. Substrates that were represented by several different synonyms (e.g. uroporphyrinogen-III = uroporphyrinogen III). were replaced with one unique name. (26 substrates out of 1316)

2. Substrates without defined chemical identity, such as "acceptor", were removed from the analysis.

Construction of the metabolic network:

After correcting database, we constructed the network for each organisms using the following steps:

1. Each pathways contains several reactions (in example shown in Fig. M1, we have 6 reactions, R1, R2, …, R6) which is composed by substrates and enzymes connected by directed links. For each reaction, educts and products are considered as nodes (and we assign a unique ID to each of them, such as S1, S2, S3…). The nodes are connected to the temporary educt-educt complexes which are also unique for specific reactions, denoted by M1, M2, …. To each temporary complex we associate an enzyme, denoted by E1, E2,…. For example, if we have the following reaction,

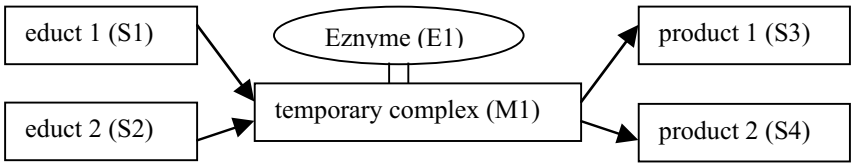



Figure M2.

We obtain the following connectivity information

S1 → [M1-E1], S2 → [M1-E1], [M1-E1]→ S3, [M1-E1]→ S4.

Bi-directional reactions (R1, R2, R3, R4, R5 in above example), were considered as two separate reactions in each direction.

Naturally, the same substrates can participate in multiple reactions, both as products and educts. For a given organism, that has $N$ substrates, $E$ enzymes and $R$ intermediate complexes, the full stochiometric interactions about the metabolic network can be compiled in an $(N+E+R)\times(N+E+R)$ matrix. For example, should an organism posseses only the reaction described in Fig. M2, the adjacency matrix would have the form:

| | S1 | S2 | S3 | S4 | M1 | E1 |
|---|---|---|---|---|---|---|
| S1 | 0 | 0 | 0 | 0 | 1 | 0 |
| S2 | 0 | 0 | 0 | 0 | 1 | 0 |
| S3 | 0 | 0 | 0 | 0 | 0 | 0 |
| S4 | 0 | 0 | 0 | 0 | 0 | 0 |
| M1 | 0 | 0 | 1 | 1 | 0 | 1 |
| E1 | 0 | 0 | 0 | 0 | 1 | 0 |

For the more complex reactions given in Fig. 1e in the manuscript we have the adjacency matrix

| | A | B | C | D | F | G | H | I | J | K | L | N | M1 | M2 | M3 | M4 | E1 | E2 | E3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| H | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| J | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| L | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| N | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| M2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| M3 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| M4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| E2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| E3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

The adjacency matrix, **A**, then contains the full connectivity information of the system, and from it one can reconstruct the full metabolic network. We generated such a matrix for each of the 43 organisms separately.

## II. Database analysis.

**Once *A* was obtained, several quantities, which are frequently used in graph theory, are measured.**

1. Connectivity distribution, *P(k)* (see ***Fig. 2***)

*$P(k_{in})$* : *Connectivity distribution for incoming links*.

Substrates generated by a biochemical reaction are products, and will be characterized by links pointing to them, i.e. incoming links. Thus, if a substrate is generated by a single reaction it will have only one incoming link, i.e. $k_{in}$ =1. For each substrate we have determined $k_{in}$ separately and prepared a histogram for each organisms, providing how many substrates have exactly $k_{in}$ =0,1,… incoming links. Dividing each element of this histogram with the total number of

substrates in a given organisms gives $P(k_{in})$, or the probability that a substrate has $k_{in}$ incoming links. This probability is shown in Fig. 2a-c for three different organisms. For Fig. 2d we have first determined $P(k_{in})$ for each organism separately, then averaged over the obtained curves. Each figure shows $P(k_{in})$ on a log-log scale, that allows us to visualize that $P(k_{in})$ follows a power law $P(k_{in}) \sim k_{in}^{-\gamma_{in}}$. A least square fit to the curve, providing the slope $\gamma_{in}$, is reported for each organism in Table 1 in column $\gamma_{in}$.

$P(k_{out})$ : *Connectivity distribution for outgoing links.*
Substrates that participate as educts in a reaction, will have outgoing links. We have performed the same analysis as above for $k_{in}$, determining the number of outgoing links ($k_{out}$) for each substrate. The results are shown as squares in Fig. 2a-c, and 2d, following the same procedure as for $k_{in}$. To reduce noise, in Fig. 2 we used a standard method applied to power law tails, called logarithmic binning i.e., in determining the histogram the bin size increases as a power of $k$. This method reduces the error bars while leaves the nature of the distribution unaffected.

2. Histogram of biochemical pathway lengths, $\Pi(l)$ (see **Fig. 3a**)

$\Pi(l)$ : For all pairs of substrates we have determined the shortest biochemical pathway, i.e., the smallest number of reactions by which starting from substrate A substrate B can be reached. For this we use a burning algorithm (in computer science is called a breadth-first-search algorithm): Starting from substrate A, we follow its outgoing links to go to the intermediate state (M) and from M, we again follow the outgoing links of M to get to the products of that reaction. Substrates (S1, S2, …,S$k$) reached in this step will be $l$=1 away from A. We continue this procedure following all outgoing links from (S1, S2, …, S$k$), the next set of substrates being $l$=2 away from A (we make sure that substrates that have been already reached are not visited again). We continue until all substrates have been reached in the network, finding the distance ($l_1$, $l_2$, …, $l_N$) between substrate A and all substrates that can be reached from A. Note that since the metabolic network is directed, i.e. A→B does not necessarily imply B→A, it is not guaranteed that there is a path between A and all other substrates. We repeat the same procedure for all substrates as a starting point of the burning algorithm, and we prepare a histogram of the obtained path lengths.

3. Diameter, $D$ (see **Fig. 3b**)

$D$ : Once $\Pi(l)$ is determined, we calculate the diameter using $D = \sum_l l \cdot \Pi(l) \Big/ \sum_l \Pi(l)$, which represents the average path length between any two substrates. To measure the diameter of the network we considered only the largest cluster, ignoring small isolated clusters (which represent less than 10% of the total number of substrates).

4. Average number of incoming (outgoing) links per node, $L/N$ (**Fig. 3c(d)**)

$L/N$ : We divide the total number of incoming (outgoing) links in the network ($L$, see table 1) and divide by total number of substrates ($N$, see table 1).

5. Substrate ranking, $r$ (**Fig. 3f**)

$\langle r \rangle_o$, $\sigma(r)$: We first identified the substrates which are present in all 43 organisms (51 substrates). We then ranked each substrate based on the number of links they had in each organisms, considering incoming and outgoing links separately. Thus we assigned rank $r$=1 for the most connected substrate with the largest number of connections, and $r$=2 for next most connected one, and so on. Thus, for each substrate a well-defined $r$ value in each organism have been defined. We next determined the average rank $\langle r \rangle_o$ for each substrate, by averaging for a given substrate the $r$ value in each of the 43 organisms. We also determined the standard deviation, $\sigma(r) = \langle r^2 \rangle_o - \langle r \rangle_o^2$. With these established values, we drew ($\langle r \rangle_o$, $\sigma(r)$) for the 51 substrates.

6. Numerical values (see **Table 1**)

$N$ : Total number of substrates that appear as an educt or product in a metabolic network for each organisms, determined from the adjacency matrix **A**.

$L$(IN/OUT) : Total number of (incoming/outgoing) links that exist in a network for each organisms, again determined from **A**.

$R$ : Total number of individual reactions or temporary intermediate states (substrate-enzyme complex).

$E$ : Total number of enzymes present in each organism.

$\gamma_{in(out)}$ : The connectivity exponent from the slope of $P(k_{in(out)})$ on a log-log plot.

$D$ : Diameter of network.

Hub(IN/OUT) : List of ten substrates with the largest number of (incoming/outgoing) links.

## III. 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 is only partially sequenced. Therefore two major sources of possible errors in the database could potentially 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. To determine if these limitations affect our analysis we have performed simulations according to the type of errors to quantitate the effect of database errors on the validity of our findings.

(a) The usefulness of integrated pathway-genome databases, such as the WIT database, relies strongly on the accurate functional annotation of a genome. Although the frequency of incorrect functional annotations in the sequence databases has not been firmly established, a recent study by Brenner [Trends Genet. 15: 132-133 (1999)] estimates the error rate to be minimally ~8% in fully sequenced microbial genomes.

A substrate that is incorrectly annotated (i.e. indicated to participate in the wrong reaction), -in "network language"- creates a wiring error, i.e. it appears as a link connected to the wrong node. Such errors typically do not modify the number of nodes or links, but randomly rewire the links in the system. An important property of scale-free networks is that such random rewiring does not change the scale-free nature of the network. To demonstrate that this is indeed the case for metabolic networks, in Fig. M3a we show the connectivity distribution of the metabolic network of *E. coli*, while in Fig. M3b and c we show the same distribution *P(k)* after 10% and 20% of the links are rewired randomly. One can see that the obtained *P(k)* is not sensitive to this type of errors. This is best seen in Fig. M3d, where we overlapped Figs. M3a-c, indicating that indeed there is no significant change in the shape of *P(k)* as a result of this rewiring process. Furthermore, Fig. M4 indicates that the diameter is essentially unchanged for as high rewiring rates as 25%. These results indicate that the estimated 8% annotation error rate would not affect the results reported in the manuscript.

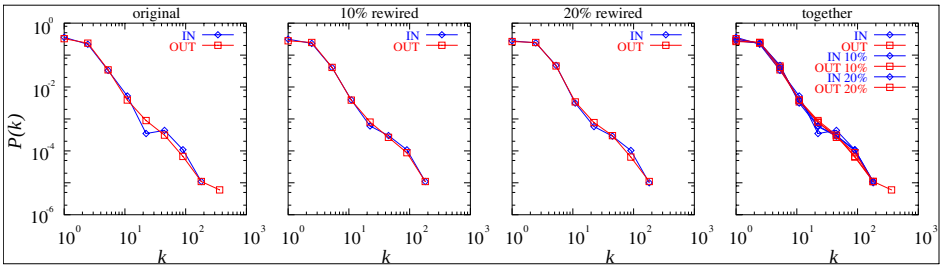



Figure M3. The effect of erroneous annotation on the connectivity distribution. (a) *P(k)* for *E. coli* , *P(k)* after (b) 10% and (c) 20% of the links have been randomly rewired for the *E. coli* network. (d) Fig. a-c superposed, demonstrating that erroneous annotation does not change *P(k)*.

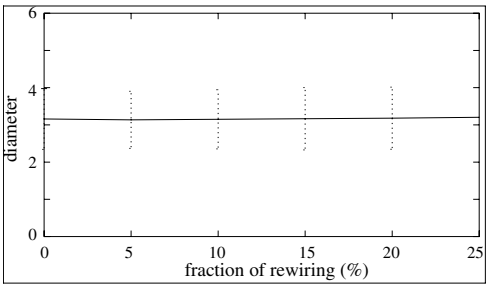



Figure M4. The effect of erroneous annotation on the diameter of the *E. coli* bacteria, indicating that as high as 25%, rewiring error creates only an insignificant changes in *D*.

(b) A more common error is the absence of reactions and pathways, either because they have not been discovered yet, or are simply omitted from the database. This is less of a problem for extensively studied, fully sequenced organisms, such as *E. coli*, but could very well apply to those organisms that have not yet been fully sequenced, such as *S. pneumoniae*. To test if this type of error would significantly affect our results we can carry out an inverse study to address the effect of missing substrates. Since the metabolic network of *E. coli* is thought to represent a network with fewest errors, we will start from it, and remove a certain fraction of the nodes randomly, mimicking substrates that for some reason are missing from the database. Fig. M5a again shows *P(k)* for the complete *E. coli*, while Figs. M5b and c show *P(k)* after 10% and 20% of the nodes have been eliminated randomly. We can observe that the connectivity distribution remains unchanged for the incomplete network, best seen in Fig. M5d, where we show the data of Fig. M5a-c together. As we have already shown in Fig. 3e of the manuscript, the diameter under such random elimination of nodes also remains unchanged. If indeed certain pathways are missing, they will likely to involve not randomly selected substrates, but those that have only a few specific connections, since substrates that participate in many reactions, with very high probability, have already been discovered and characterized. Thus the effect of missing substrates will be even less noticeable than that offered by their random removal, since with high probability only the least connected substrates are those that are missing.

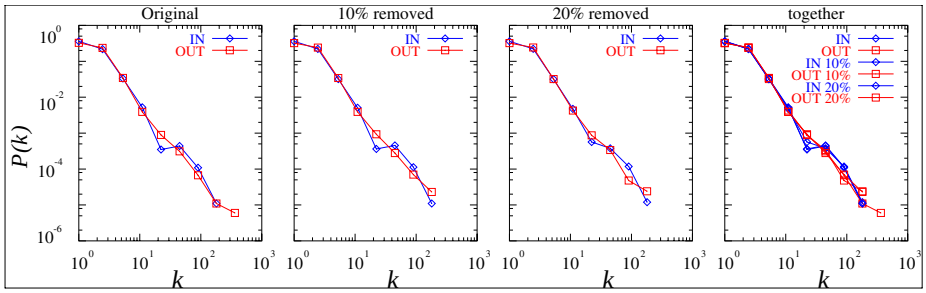



Figure M5. The effect of the absence of reactions/substrates on the connectivity distribution. In (a) we show *P(k)* for the full E. coli, and in (b)*P(k)* after 10%, (c) 20% of the nodes have been randomly removed. (d) The result (a)-(c) overlapped.

To further examine this point we show the summary *P(k)* for the completely sequenced (25) vs. incompletely sequenced (18) organisms. As one can see there is no difference between the two averages (Fig. M6).

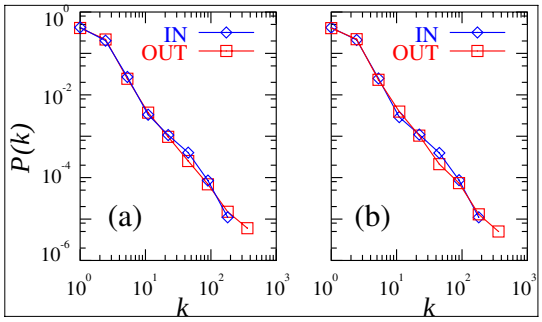



Figure M6. Averaged *P(k)* for (a) fully sequenced organisms and (b) incompletely sequenced organisms. [The incompletely sequenced organisms are: Archae: *P. horikoshii*; Bacteria: *P. gingivalis, M. bovis, M. leprae, E. faecalis, C. acetobutylicum, S. pneumoniae, S. pyogenes, C. tepidum, R. capsulatus, N gonorrhoeae, s. typhi, Y. pestis, A. actinomycetemcomitans, P. aeruginosa,* Eukaryota: *E. nidulans, O. sativa, A. thaliana.*]