The large-scale organization of metabolic networks

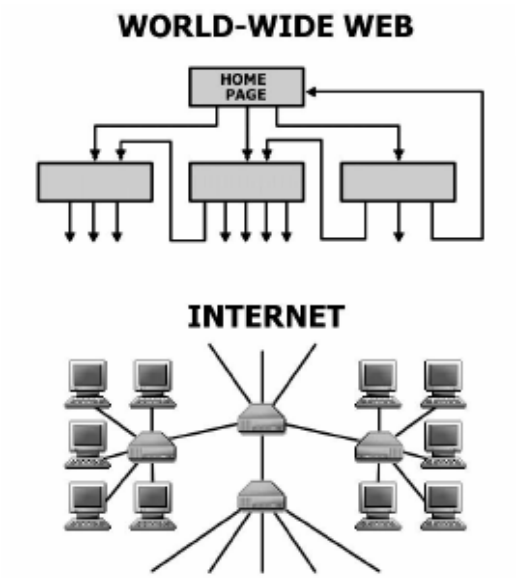
Jeong H., Tombor B., Albert R., Oltvai Z.N., and Barabási A.-L. *The large-scale organization of metabolic networks*. Nature 407, 651-654 (2000).

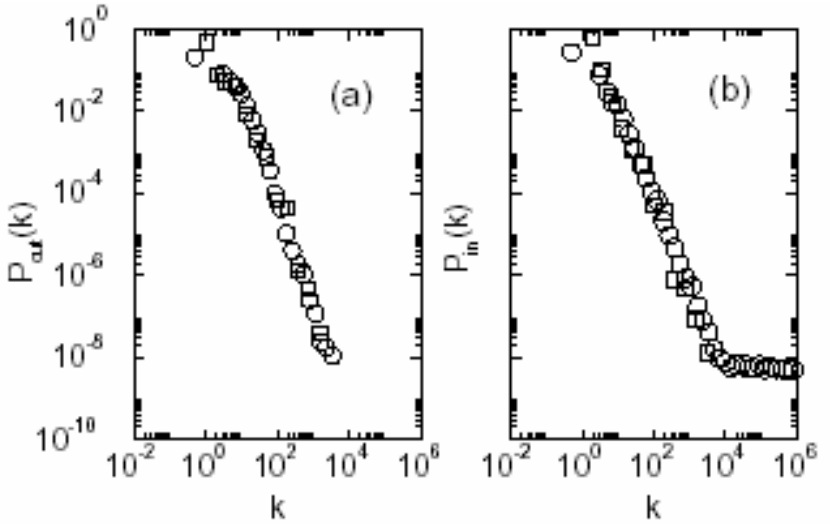
Introduction

- cellular processes → a complex network of cellular constituents and reactions
- the large-scale structure is essentially unknown
- a comparative mathematical analysis of the metabolic networks of the core metabolic network of 43 different organisms
- 18 of the 43 genomes deposited in the database are not yet fully sequenced

Empirical study indicated that WWW, Internet, Social networks → deviate from random structure
Many **real networks** show scale-free behavior

- World-Wide Web
- Internet
- Ecology network
- Science collaboration network
- Movie actor collaboration network
- Cellular network
- Phone-call network
- Network in linguistic
- Power and neural network
- Human sexual contacts
- Protein folding – conformations of a 2D lattice polymer





Real networks analysis articles

TECHNICAL COMMENT

Power-Law Distribution of the World Wide Web

Barabási and Albert (1) propose an improved version of the Erdős-Rényi (ER) theory of random networks to account for the scaling properties of a number of systems, including the link structure of the World Wide Web (WWW). The theory they present, however, is inconsistent with empirically observed properties of the Web link structure.

Barabási and Albert write that because "of the preferential attachment, a vertex that acquires more connections than another one will increase its connectivity at a higher rate; thus, an initial difference in the connectivity between two vertices will increase further as the network grows. . . . Thus older . . . vertices increase their connectivity at the expense of the younger . . . ones, leading over time to some vertices that are highly connected, a 'rich-get-richer' phenomenon" [figure 2C of (1)]. It is this prediction of the Barabási-Albert (BA) model, however, that renders it unable to account for the power-law distribution of links in the WWW [figure 1B of (1)].

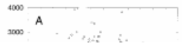
We studied a crawl of 260,000 sites, each one representing a separate domain name. We counted how many links the sites received from other sites, and found that the distribution of links followed a power law (Fig. 1A). Next, we queried the InterNIC database (using the WHOIS search tool at www.networksolutions.com) for the date on which the site was originally registered. Whereas the BA model predicts that older sites have more time to acquire links and gather links at a faster rate than newer sites, the results of our search (Fig. 1B) suggest no correlation between the age of a site and its number of links.

The absence of correlation between age and the number of links is hardly surprising; all sites are not created equal. An exciting site that appears in 1999 will soon have more links than a bland site created in 1993. The rate of acquisition of new links is probably proportional to the number of links the site already has, because the more links a site has, the more visible it becomes and the more new links it will get. (There should, however, be an additional proportionality factor, or growth rate, that varies from site to site.)

Our recently proposed theory (2), which accounts for the power-law distribution in the

data, we can illustrate the same procedure for the network of movie actors that we discussed (1). When the connectivity of the individual actors is plotted as a function of the release year of their first movie (Fig. 1A), the results are very similar to those shown in fig. 1B of Adamic and Huberman's comment. The only difference is that the movie industry had its boom not 4 years ago, as did the WWW, but rather at the beginning of the century; thus, the apparently structureless regime persists much longer. When the connectivity of the actors that debuted in the same year is averaged, however, the average connectivity in the last 60 years increases with the actor's age, in line with the predictions of our theory, and the curve follows a power law for almost a hundred years (Fig. 1B). We expect that a similar increasing tendency would appear for the WWW data after averaging, but the length of the scaling interval would be limited by the Web's comparatively brief history.

The fluctuations that lead to the apparent randomness of Fig. 1A are due to the individual differences in the rate at which nodes increase their connectivity. It is easy to include such differences in the model and continuum theory proposed by



Physica A 281 (2000) 69–77
PHYSICA A
www.elsevier.com/locate/physa

Scale-free characteristics of random networks: the topology of the world-wide web

Albert-László Barabási*, Réka Albert, Hawoong Jeong
Department of Physics, College of Science, 225 Nieuwland Science Hall, University of Notre-Dame, Notre-Dame, IN 46556, USA

Abstract

Two degrees of separation in complex food webs

Richard J. Williams*, Eric L. Berlow†, Jennifer A. Dunne*‡, Albert-László Barabási§, and Neo D. Martinez*¶

*Romberg Tiburon Center and Department of Biology, San Francisco State University, 3150 Paradise Drive, Tiburon, CA 94920; †White Mountain Research Station and Department of Integrative Biology, University of California, Berkeley, CA 94720-3140; ‡Santa Fe Institute, Santa Fe, NM 87501; and §Department of Physics, University of Notre Dame, Notre Dame, IN 46556

Communicated by Thomas W. Schoener, University of California, Davis, CA, July 26, 2002 (received for review February 12, 2002)

Feeding relationships can cause invasions, extirpations, and population fluctuations of a species to dramatically affect other species within a variety of natural habitats. Empirical evidence suggests that such strong effects rarely propagate through food webs more than three links away from the initial perturbation. However, the size of these spheres of potential influence within complex communities is generally unknown. Here, we show for that species within large communities from a variety of aquatic and terrestrial ecosystems are on average two links apart, with >95% of species typically within three links of each other. Species are drawn even closer as network complexity and, more unexpectedly, species richness increase. Our findings are based on seven of the largest and most complex food webs available as well as a food-web model that extends the generality of the empirical results. These results indicate that the dynamics of species within ecosystems may be more highly interconnected and that biodiversity loss and species invasions may affect more species than previously thought.

The mean distance between all nodes in a web (***D***) is perhaps the most familiar property of complex networks. For example network analyses (1). Because self-self trophic links may be dynamically important in ecosystems, we define ***d*** for the self-self species pair the same as for any other species pair. This method also allows us to include the important ecological distinction (8) between cannibals, which have $d = 1$, and other species, which have $d = 2$. Our method alters ***D*** among our webs an average of <1% compared with employing the more standard convention.

Although there are hundreds of food webs in the literature, the vast majority have been criticized for being incomplete, having too few species, and lacking a rigorous empirical base (4, 8–11). Therefore, we focused our analyses on seven of the largest, most comprehensive, and highest quality empirical food webs in the primary literature (Table 1; ref. 5). Three are from freshwater habitats: Skipwith Pond (12), Little Rock Lake (9), and Bridge Brook Lake (13). Two are from habitats at freshwater-marine interfaces: Chesapeake Bay (14) and Ythan Estuary (15). Two are from terrestrial habitats: Coachella Valley (8) and the island of St. Martin (16). Among these webs, ***D*** ranges between 1.40 and 2.71 and decreases with increasing connectance (Table 1, Fig. 1). On average, these values of ***D*** are 5% smaller than if we had not

letter

Comparable system-level organization of Archaea and Eukaryotes

J. Podani[1,2], Z.N. Oltvai[1,3], H. Jeong[4], B. Tombor[3], A.-L. Barabási[1,4] & E. Szathmáry[1,2]

Published online: 13 August 2001, DOI: 10.1038/ng708

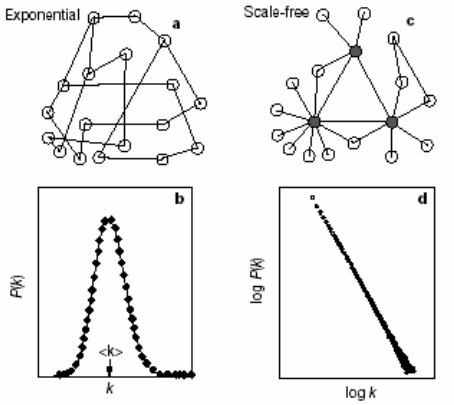
A central and long-standing issue in evolutionary theory is the origin of the biological variation upon which natural selection acts[1]. Some hypotheses suggest that evolutionary change represents an adaptation to the surrounding environment within the constraints of an organism's innate characteristics[1–3]. Elucidation of the origin and evolutionary relationship of species has been complemented by nucleotide sequence[4] and gene content[5] analyses, with profound implications for recognizing life's major domains[4]. Understanding of evolutionary relationships may be further expanded by comparing systemic higher-level organization among species. Here we employ multivariate analyses to evaluate the biochemical reaction pathways characterizing 43 species. Comparison of the information transfer pathways of Archaea and Eukaryotes indicates a close relationship between these domains. In addition, whereas eukaryotic metabolic enzymes are primarily of bacterial origin[6], the pathway-level organization of archaeal and eukaryotic metabolic networks is more closely related. Our analyses therefore suggest that during the symbiotic evolution of eukaryotes,[7–9] incorporation of bacterial metabolic enzymes into the proto-archaeal proteome was constrained by the host's pre-existing metabolic architecture.

To begin developing a systems-level understanding of the evolutionary and organizational relationships among species, we compared several characteristics of the core metabolic and information transfer pathways of 43 species from the Archaea, Bacteria, and Eukarya, based on data in the WIT integrated-pathway genome database[10]. We have previously established a graph theoretic representation of the biochemical reactions taking place in the metabolic or information transfer network of a given organism[11] (See Web Fig. A). We used the derived matrices to create four separate

Methodology

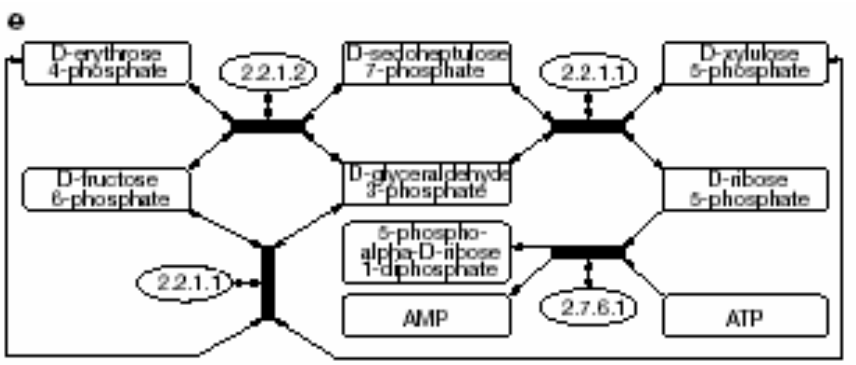
Random network (Erdos and Renyi), Scale-free network

- graphical representation of metabolic pathways
- *nodes → substrates, edges → metabolic reactions*
- *some substrates can participate in multiple reactions → multiple edges*
- *enzymes (the catalytic scaffolds) → EC numbers (2.2.1.1, 2.2.1.2 … etc.)*



Random network

Scale-free network



A portion of *E. coli's* metabolic pathway. 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). Some reactions are **reversible** while others are **irreversible**, which is denoted as **simple arrow** and **double arrow links** respectively.

Node degree of connection P(k)

In a random network, the links are randomly connected and most of the nodes have degrees close to <k>. The degree distribution P(k) vs. k is a Poission distribution, i.e. $P(k) \approx e^{-<k>} \frac{< k >^K}{k!}$, for $k << <k>$ and $k >> <k>$.

In many real life networks, the degree distribution has no well-defined peak but has a power-law distribution, $\mathbf{P(k) \sim k^{-\gamma}}$,where γ is a constant. Such networks are known as **scale-free network**.

Random network → *Log[P(k)] vs Log[k]* plot has a peak.

Scale-free network → *Log[P(k)] vs Log[k]* plot is a line with negative slope.

Results

- compute k_{in} and k_{out} for every substrate

Node degree of connection

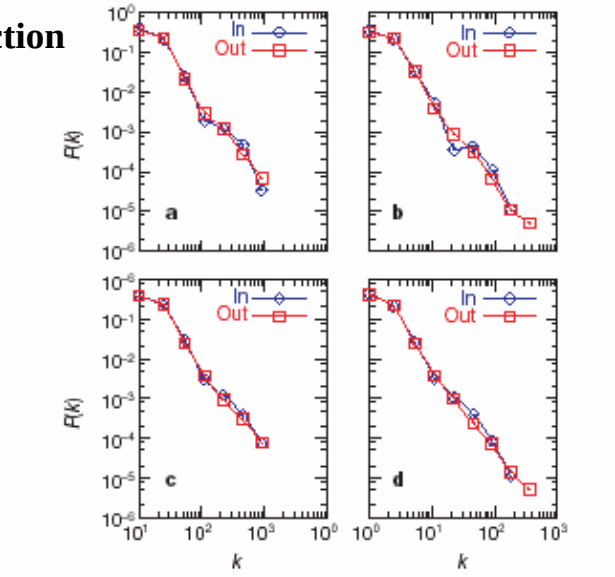


Figure 2 Connectivity distributions $P(k)$ for substrates. **a**, *Archaeoglobus fulgidus* (archae); **b**, *E. coli* (bacterium); **c**, *Caenorhabditis elegans* (eukaryote), shown on a log–log plot, counting separately the incoming (In) and outgoing links (Out) for each substrate. k_{in} (k_{out})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

- Fig. 2a,b,c ➔ **substrates participate in multiple reactions are rare** for three different species (*Archaeoglobus Fulgidus* (archae), *E. coli* (bacterium), *Caenorhabditis elegans* (eukaryote))
- scale-free network ➔ $P(k) < k^{-\gamma}$, with $\gamma in = 2.2$, $\gamma out = 2.2$
- Fig. 2d ➔ the connectivity distribution averaged over 43 organisms
- **Metabolic networks belong to the class of scale-free networks**

Properties of metabolic networks

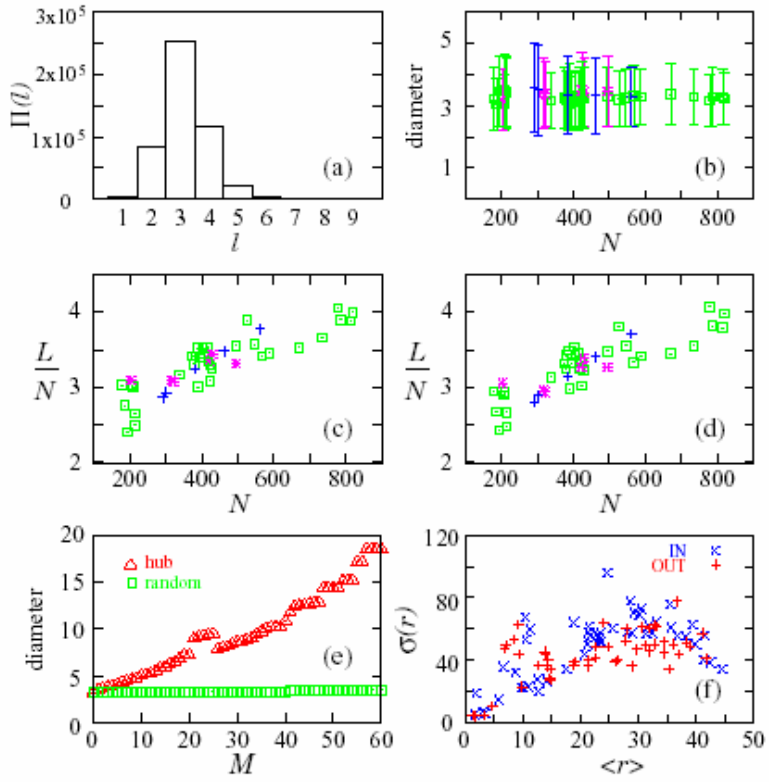


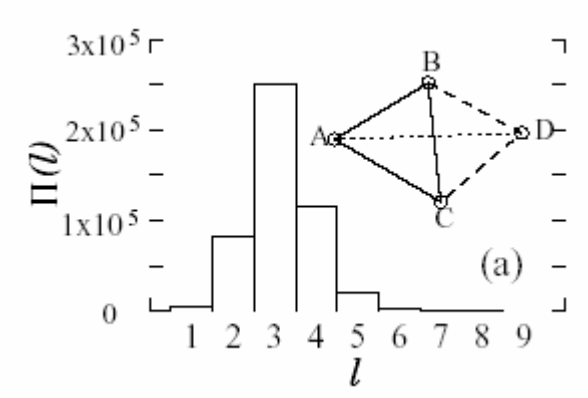
Figure 3

(**a**) The histogram of the biochemical pathway lengths, ℓ, in the bacterium, *E. coli*. (**b**) The average path length (diameter) for each of the 43 investigated organisms. The error bars 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*). (**c**) The average number of incoming links or (**d**) outgoing links per node for each organism. (**e**) The effect of substrate removal on the metabolic network diameter of the bacterium, *E. coli*. In the upper curve (Δ) in an inverse order of connectivity, the most connected substrates are removed first. In the bottom curve (□) 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 (σ_r) as a function of the average ranking, $\langle r \rangle_o$ for substrates present in all 43 investigated organisms. The horizontal axis in (b,c,d,) denotes the number of nodes in each organism. Archaea (magenta), bacteria (green), and eukaryotes (blue) are shown in (b,c,d,f).

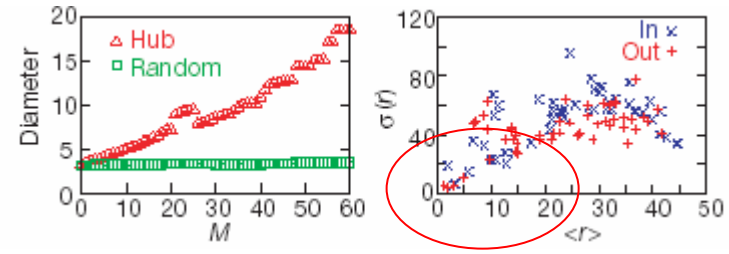
Constant diameter

- Fig. 3b → the **average path length (diameter) is constant for each of the 43 organisms** !!
- For non-biological network, network diameter is proportional to total number of nodes → $d \sim \log N$
- Fig. 3c, d → 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.
- a possible advantage of small mean path lengths stems from the importance of minimizing transition times between metabolic states in response to environmental changes (Easterby 1986; Schuster and Heinrich 1987; Cascante et al. 1995). Networks with robustly small average path lengths thus might adjust more rapidly to environmental change.
-

An unchanged diameter can be maintained through increasing the average connectivity



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 $D=(l_{AB}+l_{AC}+l_{BC}+l_{BD}+l_{DC}+l_{AD})/6 = 7/6$ since $l_{AD}=2$ 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**.



Metabolic networks are robust against random errors

- Fig. 3e → **removal of the most connected substrates the diameter increases rapidly**, illustrating the **special role of these metabolites** in maintaining a constant metabolic network diameter.
- when **a randomly chosen M substrates are removed** (~8%) – mimicking the consequence of **random mutations of catalyzing enzymes** – the **average distance between the remaining nodes is not affected**, indicating a striking **insensitivity to random errors**.
- 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**.

Study of highly connected substrates

- same substrates act as the hubs in all organisms ? or organism-specific of the most connected substrates
- the ranking of the most connected substrates is **practically identical for all 43 organisms** → **generic utilization** of the same substrates by each species
- the standard deviation σ_r **increases with the average rank order** $<r>$, 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

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	dabcgifeol
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 **γin and γout** 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=$NH4^+$, m=CoA, n=AMP, o=pyruvate, p=L-glutamine, q=2-oxoglutarate, r='alpha'-Dglucose1-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.

- Metabolic networks have a history.
- How does a network arrive at a power-law degree distribution if it grows?
- First and unsurprisingly, it adds nodes to a graph.
- Second, it connects this node to previously existing nodes according to a second rule, where already highly connected nodes are more likely to receive a new connection than nodes of lesser connectivity. Over many node additions, a power law degree distribution emerges.
- Highly connected nodes are old nodes, nodes having been added very early in a network's history.
- *Are highly connected metabolites old metabolites?*
-

Twelve key metabolites in *E. coli* ranked by degree ("connectivity")

glutamate (51)	**Y**
pyruvate (29)	**Y**
coenzyme A (29)	**Y**
α-ketoglutarate (27)	**?**
glutamine (22)	**Y**
aspartate (20)	**?**
acetyl-CoA (17)	**Y**
phosphoribosyl pyrophosphate (16)	**?**
tetrahydrofolate (15)	**?**
succinate (14)	**?**
3-phosphoglycerate (13)	**?**
serine (13)	**?**

- ☐ **proposed remnants of a surface metabolism or an RNA world**
- ☐ **proposed early amino acids**
- ☐ **proposed early metabolites (in the tricarboxylic acid cycle or in glycolysis)**

The network was generated after the elimination of the compounds ***NAD***, ***ATP***, and ***their derivatives***. These are even more highly connected than the compounds shown here. They are also evolutionarily ancient.

5 out of 12 are found in Table 1

Warning

- the chemical reaction networks of planetary atmospheres, networks largely shaped by the photochemistry of their component substrates
- **chemical reaction networks in these atmospheres**, despite being vastly different in chemistry, h**ave a degree distribution consistent with a power law** (Gleiss et al. 2001). This suggests that power-law distributions may be very general features of chemical reaction networks. The reasons why we observe them in **cellular reaction networks may have nothing to do with the robustness they may provide**

References

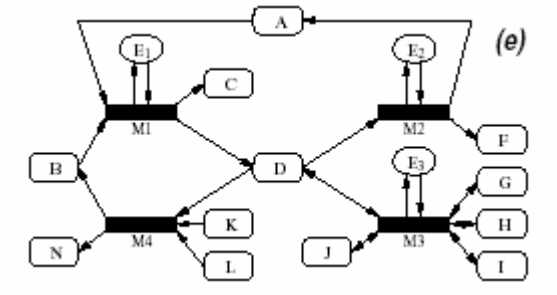
Cascante, M., E. Melendezhevia, B. Kholodenko, J. Sicilia, and H. Kacser. 1995. Control Analysis of Transit-Time For Free and Enzyme-Bound Metabolites : Physiological and Evolutionary Significance of Metabolic Response-Times. *Biochemical Journal* **308**: 895-899.

Easterby, J.S. 1986. The Effect of Feedback On Pathway Transient-Response. *Biochemical Journal* **233**: 871-875.

Gleiss, P.M., P.F. Stadler, A. Wagner, and D.A. Fell. 2001. Small cycles in small worlds. *Advances in complex systems* **4**: 207-226.

Schuster, S. and R. Heinrich. 1987. Time Hierarchy in Enzymatic-Reaction Chains Resulting From Optimality Principles. *Journal of Theoretical Biology* **129**: 189-209.

Exercise



A portion of the metabolic pathway of an unknown organism is depicted above. Determine

(1) the in and out degree of nodes A, B, C, D, F and G, and

(2) the intermediate reactions and path length of the reactions.

Node	k_{in}	k_{out}
A		
B		
C		
D		
F		
G		
Link	Intermediate reactions	Path length
B – D	$B \xrightarrow{E_1} D$	1
B – F	$B \xrightarrow{E_1} D \xrightarrow{E_2} F$	2
A – B		
A – C		
A – F		