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)

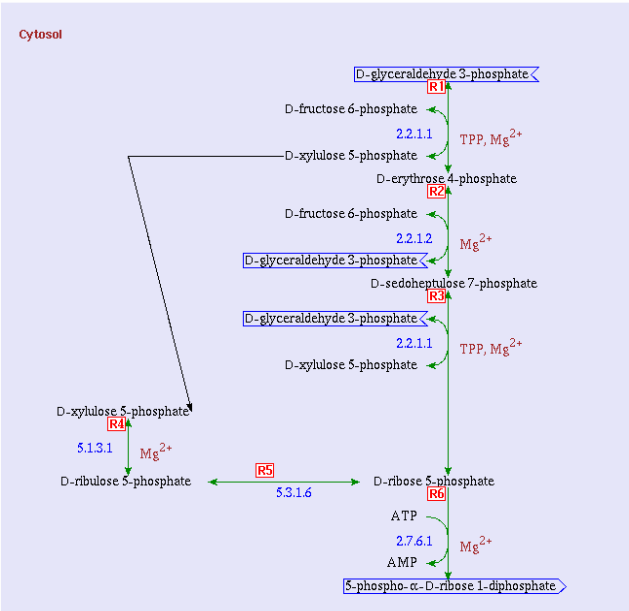

Figure M1.

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,

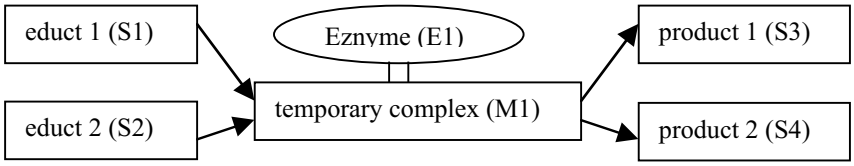

Figure M2.

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 γ_{in}, is reported for each organism in Table 1 in column γ_{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, …,Sk) reached in this step will be l=1 away from A. We continue this procedure following all outgoing links from (S1, S2, …, Sk), 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.

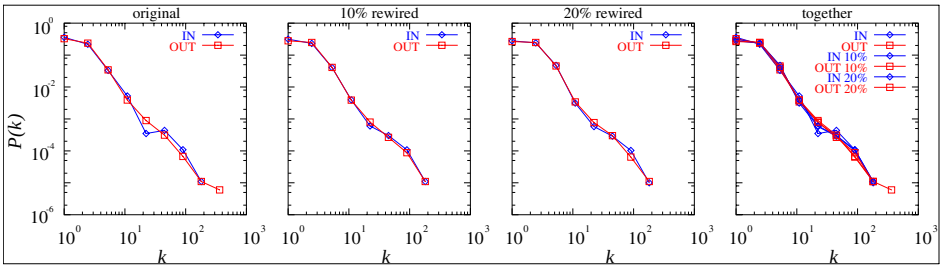

Figure M3. The effect of erroneous annotation on the connectivity distribution. (a) *P(k)* for *E. coli* , *P(k)* after (b) 10% and (c) 20% of the links have been randomly rewired for the *E. coli* network. (d) Fig. a-c superposed, demonstrating that erroneous annotation does not change *P(k)*.

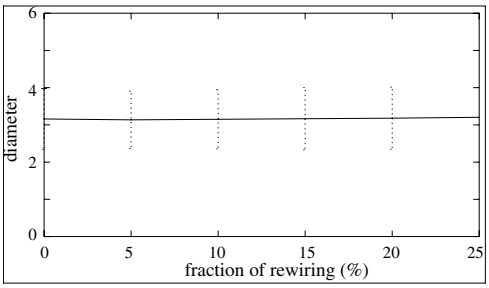

Figure M4. The effect of erroneous annotation on the diameter of the *E. coli* bacteria, indicating that as high as 25%, rewiring error creates only an insignificant changes in *D*.

(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.

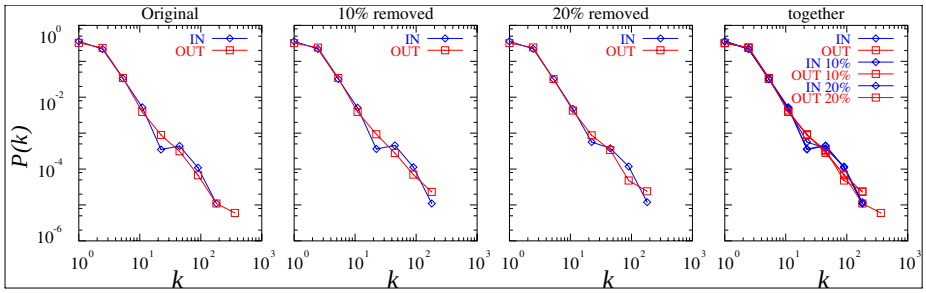

Figure M5. The effect of the absence of reactions/substrates on the connectivity distribution. In (a) we show *P(k)* for the full E. coli, and in (b)*P(k)* after 10%, (c) 20% of the nodes have been randomly removed. (d) The result (a)-(c) overlapped.

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).

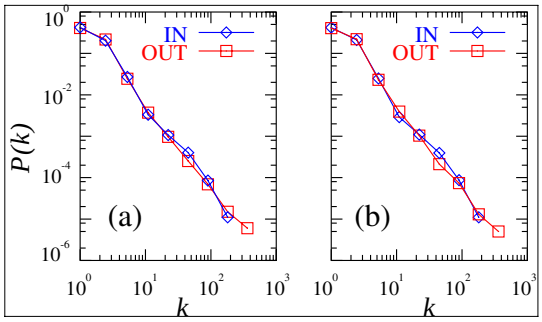

Figure M6. Averaged *P(k)* for (a) fully sequenced organisms and (b) incompletely sequenced organisms. [The incompletely sequenced organisms are: Archae: *P. horikoshii*; Bacteria: *P. gingivalis, M. bovis, M. leprae, E. faecalis, C. acetobutylicum, S. pneumoniae, S. pyogenes, C. tepidum, R. capsulatus,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