Supplementary Information

For the article**"*Comparable system-level organization of Archaea and Eukaryotes*"** by J. Podani, Z. N. Oltvai, H. Jeong, B. Tombor, A.-L. Barabási, and E. Szathmáry

(reference numbers are the same as in the paper)

Web Note A

Connectivity distribution of information transfer networks

To determine the large-scale structural organization of information transfer networks, we have examined the topologic properties of the information transfer pathways of 43 different organisms based on "Information transfer" portions of data deposited in the WIT database[10]. Similar to that of metabolic networks[11], we have first established a graph theoretic representation of the biochemical reactions taking place in a given information transfer network. In this representation, a network is built up of nodes, which are the substrates that are connected to one another through links, which are the actual biochemical reactions. Note, that beside core metabolic reactions, the WIT DB includes biochemical reactions contributing to (1) information pathways, (2) electron transport, (3) transmembrane transport, (4) signal transduction, and (5) structure and function of the cell. However, with the exception of information transfer pathways, the data deposited for these parts are so incomplete that comparison of e.g., signal transduction pathways is completely impossible. In addition, no information for any metazoan, including for those of the just recently sequenced human genome, is available in the WIT DB.

To establish if the topology of information transfer networks is best described by the inherently random and uniform exponential- or the highly heterogeneous scale-free model[11] the connectivity distribution of all 43 networks were analyzed as described for metabolic networks in ref. 11. As illustrated in Web Fig. A, our results convincingly indicate that the probability that a given substrate participates in *k* reactions follows a power-law distribution, i.e., information transfer networks belong to the class of scale-free networks. Since under physiological conditions a large number of biochemical reactions (links) in a network are preferentially catalyzed in one direction (i.e., the links are directed), for each node we distinguish between incoming and outgoing links. For instance, in *Escherichia coli* the probability that a substrate participates as an educt in *k* information transfer pathways follows $P(k) \sim k^{-\gamma_{in}}$, with $\gamma_{in} = 2.1$, and the probability that a given substrate is produced by *k* different information transfer reactions follows a similar distribution, with $\gamma_{out} = 2.2$ (Web Fig. A*b*). We find that scale-free networks describe the information transfer reaction networks in all organisms in all three domains of life (Web Fig. A*a*-*c*), indicating the generic nature of this structural organization (Web Fig. A*d*).

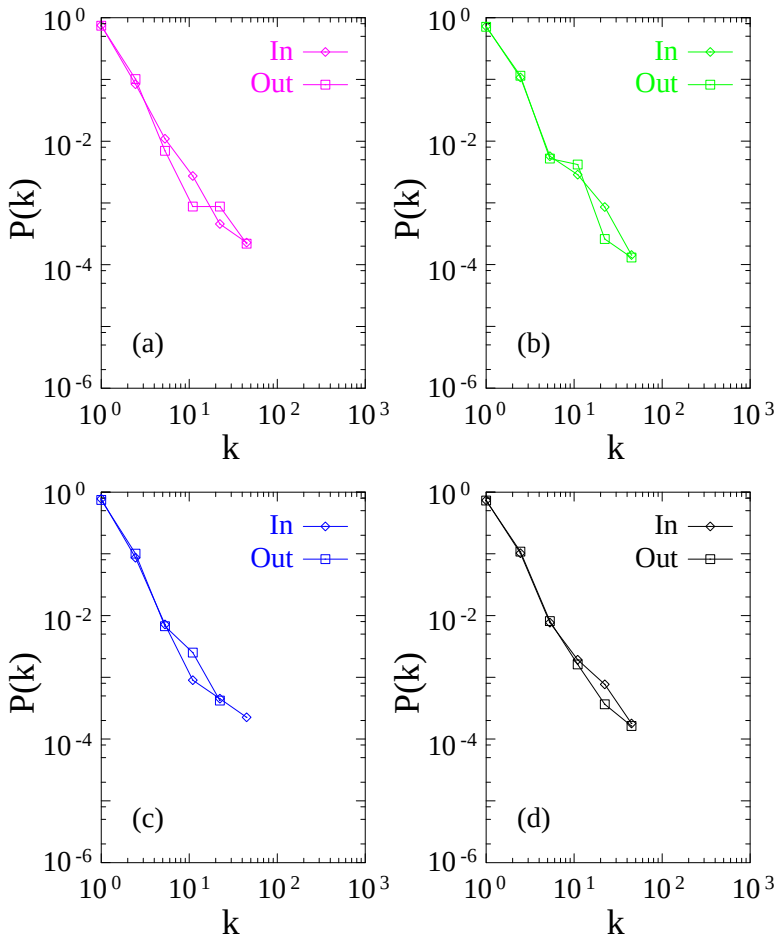

Fig. A Connectivity distribution *P(k)* for the substrates in (***a***) *A. fulgidus* (Archaea) (***b***) *E. coli* (Bacteria) (***c***) *C. elegans* (Eukarya), shown on a log-log plot, counting separately the incoming (IN) and outgoing links (OUT) for each substrate, k_{in} (k_{out}) corresponding to the number of reactions in which a substrate participates as a product (educt). (***d***) The connectivity distribution averaged over all 43 organisms.

Web Note B

Database analysis

Data types and resemblance coefficients

Although the data matrices comprise quantitative information, we decided to rely only upon the qualitative part of the data, for several reasons. First of all, traditionally, qualitative data have been almost exclusively used in molecular systematics. Second, qualitative data are less sensitive to the different intensity by which the organisms are studied, meaning that there is more noise in quantitative data attributable to 'sampling error'. The *presence or absence* (P/A) of a variable in a given organism is the simplest type of all data. We thus compared each pair of organisms using a simple ratio in which the number of variables present in both was divided by the number of variables present in at least one of them (widely known as the *Jaccard index*). The coefficient is expressed here as its complement to measure the dissimilarity of the two species being compared. In addition to the P/A scores, we also utilized the *ordinal information* (i.e., the rank order) in the data. We have previously demonstrated[11] that ordering relationships for each organism are also of great importance when describing metabolic pathways. Consequently, we also compared each pair of species using the ordinal measure suggested by Goodman and Kruskal (their γ function). In this coefficient, the number of variable pairs ordered similarly by both taxa is divided by the total number of pairs of variables that are ordered at all. For taxa *j* and *k* it is written as $\gamma_{jk} = (a - b) / (a + b)$ where *a* is the number of pairs of variables ordered for objects *j* and *k* identically, *b* is the number of pairs of variables that are reversely ordered in *j* and *k*. The coefficient ranges from -1 to 1, indicating complete disagreement and full agreement, respectively. To simplify calculations, we transformed the values to measure dissimilarity in the range of 0 to 2. For details of the methodology used in this paper see refs. 12-15.

Ordination

A major group of exploratory data analysis tools aims to reduce the dimensionality of the system, more precisely, to show a multidimensional picture in a few, preferably two, dimensions as efficiently as possible. These procedures are commonly referred to as ordination. For the present study, we have chosen the non-metric multidimensional scaling (NMDS) method which is best suited to input dissimilarities that are ordinal anyway. For

consistency, we applied the very same procedure to the otherwise metric Jaccard coefficient, as well. A fundamental feature of NMDS is that the user defines the final dimensionality of the result, which is generally selected to be two. The procedure then attempts to arrange the points in the space such that the ordering relationships of input dissimilarities are as faithfully preserved in the ordination space as possible. The procedure is iterative, and therefore requires several runs from which we retain the best result.

Hierarchical clustering

The most widely known hierarchical clustering method is *group average clustering* (UPGMA, ref. 13). Despite strong criticisms expressed by cladists, the method applies even to phylogenetic problems especially if the molecular clock (i.e., constant rate of evolutionary change on all lineages) is assumed**.** Nevertheless, UPGMA has been most commonly used on data other than DNA or RNA sequences, thus its properties and interpretability of its results are sufficiently known. In the present study, UPGMA was applied to matrices of the Jaccard index. The method produces an ultrametric tree, the hierarchical levels potentially interpretable as mean dissimilarities between groups fused at that level. Since the matrix calculated by Goodman-Kruskal's γ comprises non-metric information, they were processed by a different hierarchical classification algorithm, namely *ordinal hierarchical clustering* (OC, see ref. 14 for its underlying criterion).

As OC is a method recently developed by one of us (JP); here we provide a more detailed description of this technique. This procedure considers only the ordering information present in the dissimilarity matrix, and then builds a tree in which each level is labeled only by the number of the iteration step in which the actual merger took place. The algorithm of agglomerative ordinal clustering first orders the $m(m-1)/2$ dissimilarity coefficients (m is the number of taxa), so that each γ_{jk} is replaced by its rank, r_{jk}. Then, the following clustering criterion is considered

$$C = (R_w - R_{min}) / (R_{max} - R_{min})$$

where R_w is the sum of ranks of within-cluster dissimilarities, R_{min} is the possible minimum of such ranks for the given number of clusters and for the given numbers of objects in each cluster, and R_{max} is the possible maximum. A similar criterion – using R_{exp} instead of R_{max} – was proposed by ref. 14 for determining the importance of variables in classification. The value of C ranges from 0 to 1, 0 indicating that all within-cluster

dissimilarities are smaller than the between-cluster dissimilarities, whereas 1 indicating that all between-cluster dissimilarities are smaller than the others. In the dendrogram resulted from agglomerative ordinal clustering we used the ranks of fusions (values from 1 to *m*-1), rather than the *C* values themselves, because this criterion does not change monotonically. That is, the result is an *ordered dendrogram*, rather than a weighted dendrogram, being consistent with the ordinality of the previous steps of the analysis.

Neighbor joining (NJ)

The method originally developed by Saitou and Nei (ref. 12) has been the most widely applied distance-based tree-building method in phylogenetic reconstruction. Contrary to UPGMA, the tree is not ultrametric, but the additivity of branch lengths holds. In other words, the sum of branch lengths along the path between two taxa in the tree is an approximation to the input distances. The NJ tree is unrooted by definition, although some external criterion (outgroup, longest path) may be used to position a root (a practice not followed here).

Web Note C

Analysis of the effect of database errors

Of the 43 organisms whose metabolic- and information transfer pathways we have analyzed the genome of 25 has been completely sequenced (5 Archaea, 18 Bacteria, 2 Eukaryotes), while the remaining 18 are 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 repeated analyses according to the type of errors to demonstrate if database bias critically influences the validity of our findings. These analyses were restricted to the UPGMA classifications of the P/A case, assuming that ordinations and unrooted trees are similarly influenced by data changes. In the first series of simulations, the four UPGMA classifications (Fig. 1*f,h*, Fig. 2*f,h*) were recomputed based on a reduced set of input data such that only the randomly chosen 50% of original variables were retained. The INFO based dendrograms (compare original Fig. 1*f,h* to the corresponding ones in Web Fig. B) suffered little changes owing to data reduction, the only discrepancy being that Archaea and Eukaryotes form three, rather than two small clusters, while Bacteria remain intact as a very tight group. For METAB data, removal of 50% of the variables from the substrate data set caused only one critical modification of the tree (compare original Fig. 2*f* to the corresponding one in Web Fig. B); *Aeropyrum pernix* was moved to the B2 group which was otherwise split into two. On the other hand, the main groupings in the dendrogram of Fig. 2*h* (ENZ data) did not change at all. In the second series of simulations, the incompletely sequenced organisms were simply removed from the analyses, so that clustering was confined to 25 taxa in case of all the four input data sets. Reduction of sample size did not alter the main groupings at all (compare original Fig. 1*f,h* and 2*f,h* to the corresponding trees in Web Fig. C), suggesting robustness of the data upon the removal of incompletely sequenced organisms.

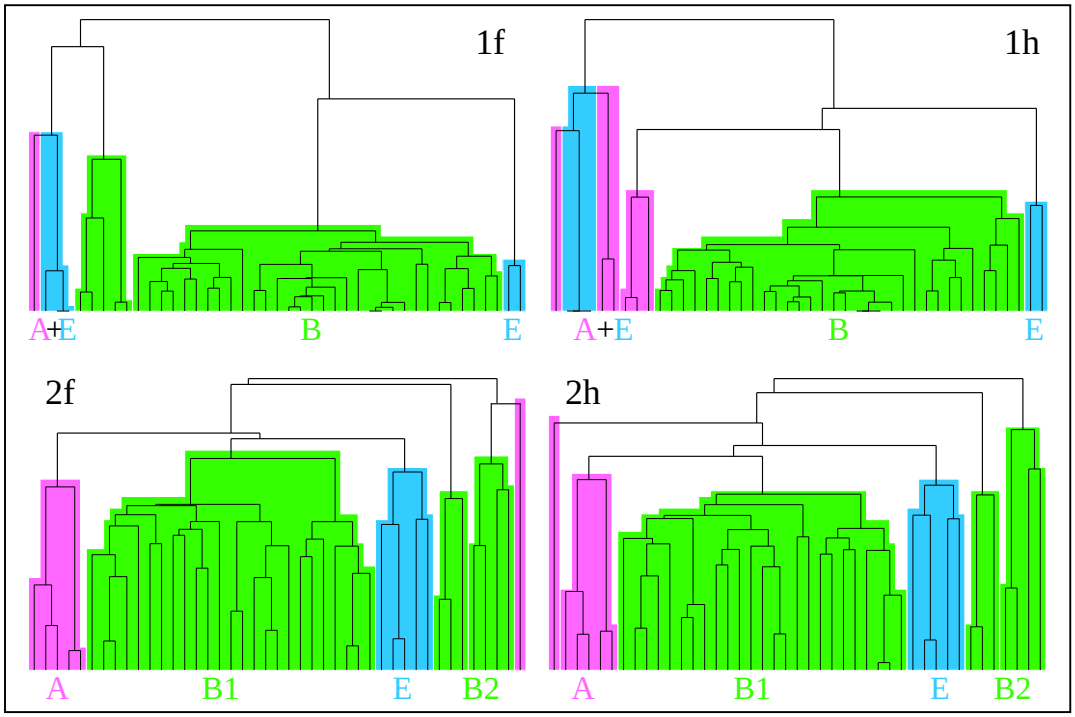

Fig. B UPGMA classifications based on randomly reduced data sets to evaluate potential effects of missing reactions and pathways. Captions refer to comparable dendrograms obtained from complete data.

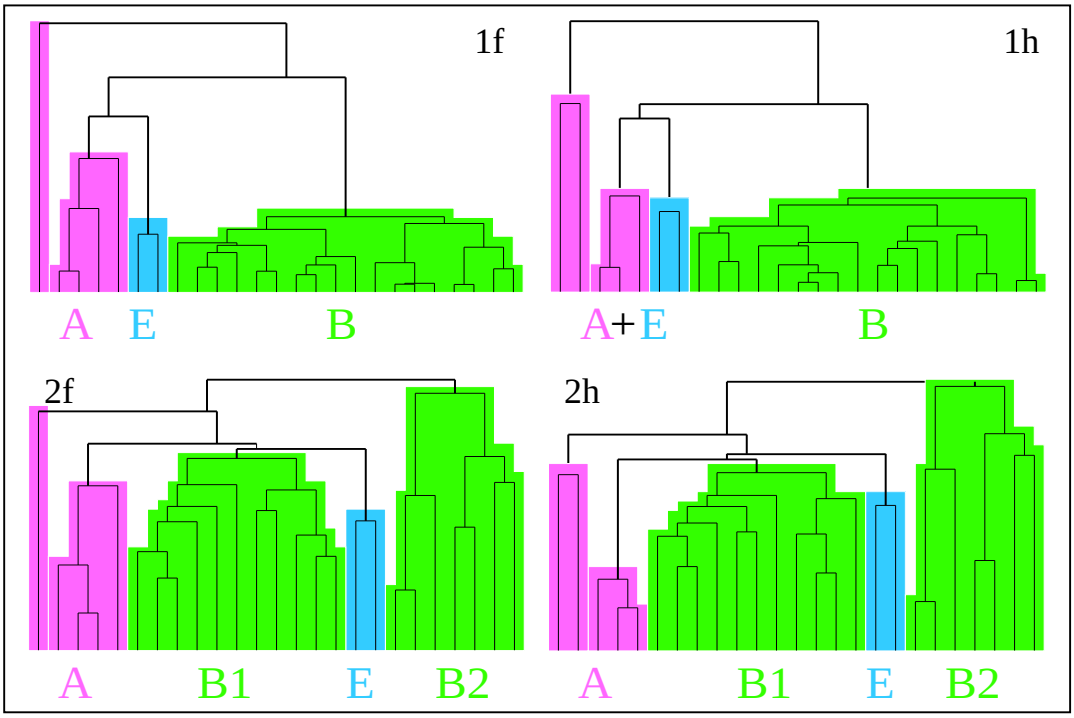

Fig. C UPGMA classifications restricted to completely sequenced species to evaluate the effect of missing information arising from incomplete sequencing. Captions refer to comparable dendrograms obtained for the full set of species.