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 data sets for each organism, comprised of substrates and enzymes of their metabolic and information transfer networks, respectively. This allowed us to determine for each data set whether a particular substrate or enzyme was present or absent in a given organism, and also to systematically rank order all the substrates and catalyzing enzymes based on the number of biochemical reactions in which they participate. We then compared the individual data sets using several different multivariate analytical approaches, including neighbor joining (NJ: the simplest distance-based cladistic method[12]), unweighted group average clustering (UPGMA: the most commonly used hierarchical classification method[13]), ordinal clustering (OC: a hierarchical method suited specifically to rank orders[14,15]) and nonmetric multidimensional scaling (NMDS: the most general ordination procedure[15]).

We first examined the information pathways (Fig. 1) by analyzing both substrates (Fig. 1*a,b,e,f,i*) and catalyzing enzymes (Fig. 1*c,d,g,h,j*) to obtain ordinations (Fig. 1*a–d*), hierarchical classifications (Fig. 1*e–h*) and unrooted trees (Fig. 1*i,j*). The results of all three approaches were highly congruent. Regardless of whether ordinal information (rank ordering; Fig. 1*a,c,e,g*) or simple presence/absence (P/A) (Fig. 1*b,d,f,h–j*) was considered, and whether the substrate or enzyme

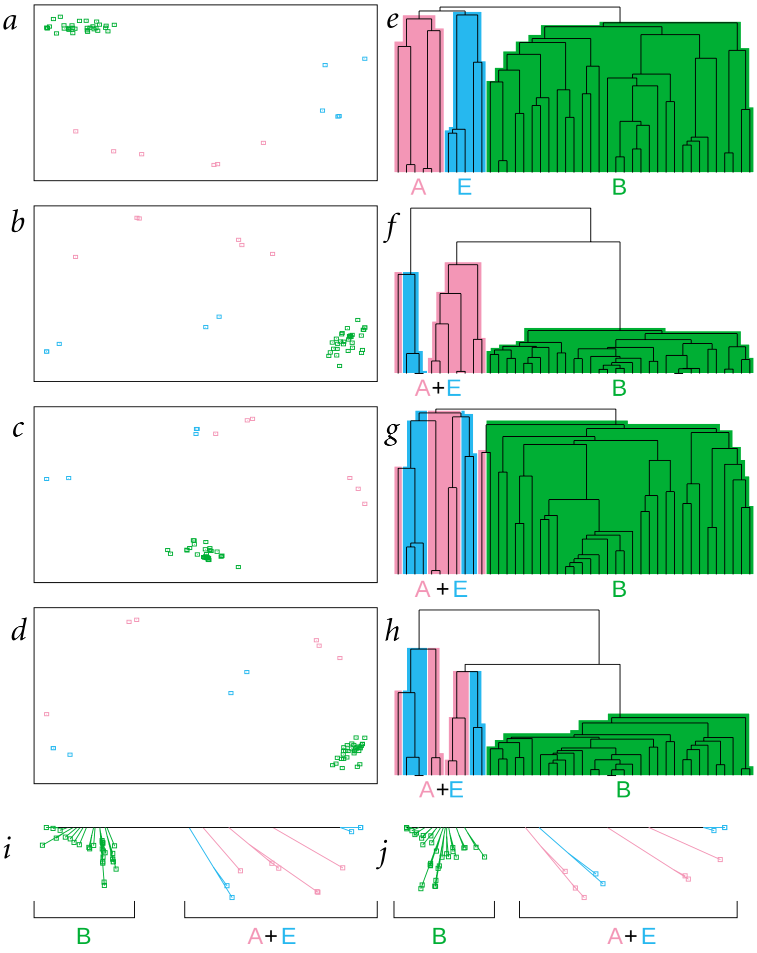

Fig. 1 Analyses based on information transfer pathways. ***a–d***, NMDS ordinations. ***e,g***, OC classifications. ***f,h***, UPGMA classifications. ***i,j***, unrooted NJ trees. (*a,b,e,f,i*) represent data based on substrate list;(*c,d,g,h,j*) are based on enzyme variables. (*a,c,e,g*) represent ordinal information; (*b,d,f,h,i,j*) represent P/A information. A, Archaea; B, Bacteria; E, Eukarya.

[1]Institute for Advanced Study, Collegium Budapest, H-1014 Budapest, Hungary. [2]Department of Plant Taxonomy and Ecology, Eötvös University, H-1117 Budapest, Hungary. [3]Department of Pathology, Northwestern University Medical School, Chicago, Illinois 60611, USA. [4]Department of Physics, University of Notre Dame, Notre Dame, Indiana 46556, USA. Correspondence should be addressed to Z.N.O. (e-mail: zno008@northwestern.edu) or E.S. (e-mail: szathmary@zeus.colbud.hu).

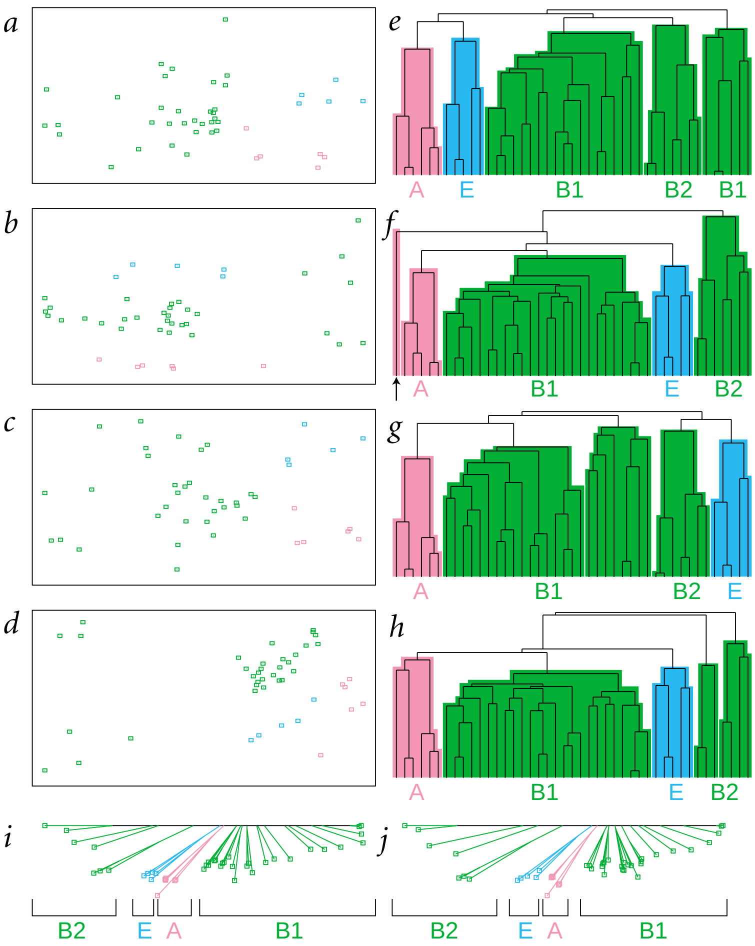

Fig. 2 Analyses based on metabolic pathways. ***a–d***, NMDS ordinations. ***e,g***, OC. ***f,h***, UPGMA classifications. ***i,j***, unrooted NJ trees. (*a,b,e,f,i*) represent data based on substrate list; (*c,d,g,h,j*) are based on enzyme variables. (*a,c,e,g*) represent ordinal information; (*b,d,f,h,i,j*) represent P/A information. A, Archaea; B, Bacteria; B1, nonparasitic bacteria; B2, parasitic bacteria; E, Eukarya. The arrow in (*f*) indicates the location of the Crenarchae *A. pernix*.

data set was used, our analyses suggest clear separation of Bacteria both from Archaea and Eukarya. This finding is in general agreement with cladistic results based on ribosomal RNA sequences[4] or gene content[5]. It is notable that despite the comparatively large number of Bacteria in the sample, the within-group differences in the Bacteria are considerably smaller than those within the Archaea or Eukara. This is manifested in the ordinations as a very compact scatter of points representing the Bacteria (Fig. 1*a*–*d*). It is also evident that Archaea and Eukarya are not merely close to each other, but in most cases are essentially inseparable. For instance, the points representing Archaea and Eukaryotes in both P/A-based ordinations, and in the ordinal case for the enzyme-based ordinations, form elongated point swarms, whose internal cohesion and segregation are therefore not supported (Fig. 1*b*–*d*). The only ordination and classification suggesting fair distinction between Archaea and Eukarya are those based on ordinal information from the substrate data (Fig. 1*a,e*).

We next compared the systemic organization of the metabolic networks. Analyses of data for the substrates (Fig. 2*a,b,e,f,i*) and the catalyzing enzymes (Fig. 2*c,d,g,h,j*) of intermediate metabolism provide a slightly more complex picture than for the informational networks, but with many similarities. In each hierarchy, Archaea and Eukarya are clearly recognizable as intact groups, with the exception of the Crenarchae *Aeropyrum pernix*, which forms a separate cluster (Fig. 2*f*). Note, however, the contentious phylogenetic classification of this organism[16]. Except in one case (Fig. 2*g*), ordinal analysis of substrate (Fig. 2*a,e*) and enzyme data (Fig. 2*c,g*) shows a clear separation of Bacteria from both Archaea and Eukarya, but a fairly close proximity between the latter two domains. For the P/A information (Fig. 2*b,d,f,h–j*), applying the NJ approach to both the substrate and enzyme data sets (Fig. 2*i,j*), and applying the NMDS approach to the enzyme data (Fig. 2*d*), support a similarly clear separation. However, using the NMDS approach with the substrate data (Fig. 2*b*) and using the UPGMA dendrograms with the substrate (Fig. 2*f*) and enzyme data (Fig. 2*h*) indicate a substantially looser association.

There is a clear separation between nonparasitic bacteria and parasitic bacteria (such as *Chlamydia pneumoniae, Mycoplasma genitalium, Rickettsia prowazekii, Treponema pallidum,* and *Borrelia burgdorferi*), which possess an evolutionarily reduced genome[17]. The parasitic bacteria almost always form one group, with the exception of one unrooted tree (Fig. 2*h*), whereas P/A analyses show the nonparasitic bacteria as a single large group in both trees. OC, however, produces two subclusters within the nonparasitic bacteria (Fig. 2*e,g*). The NJ method separates the two groups of Bacteria on the basis of P/A data for both substrates and enzymes (Fig. 2*i,j*): the parasitic bacteria form a loosely arranged group at one end of the unrooted tree, followed by Eukarya, Archaea, and finally the nonparasitic bacteria. Thus, Archaea and Eukarya are inserted between the two bacterial groups in the tree; it cannot be rooted to produce a cladogram with parasitic and nonparasitic bacteria as sister groups. It may be expected that parasitic organisms more readily lose metabolic genes, but it is notable that this loss puts them in a single statistical group, whatever their phylogenetic distance from one another.

The analysis and comparison of the systemic higher-level organization of these 43 species reveals unanticipated features of the relationships within and among the major domains. The comparison of systemic attributes of metabolic networks, for example, indicates convergent evolutionary trends that are reflected only in the metabolism of organisms. Conversely, our results also indicate that comparison of system-level features by the current methods will not allow for the identification of precise phylogenetic relationships among species, because metabolic organization is intimately tied to the environment in which they evolved. There is the additional, unexpected finding that the Archaea and Eukarya are related not only in their informational pathways but also in their metabolic pathways. Although previous sequence comparisons found the informational genes of eukaryotes to be similar to those of archaea, with respect to operational (metabolic) genes, bacteria and eukaryotes are more closely related[6,18]. Eukarya apparently diverged from Archaea[8,9,19], partly as a result of robust horizontal transfer of operational genes[8,9], or perhaps as a by-product of (imperfect) phagotrophic consumption of Bacteria[20]. Although the transfer of physically clustered, functionally complementary bacterial genes into the archaean host's genome appears to have been widespread[6,21], our analyses demonstrate that the overall eukaryotic metabolic network architecture remained significantly less changed. Thus, natural selection may have induced the differential retention of the bacterial enzymes that were presumably transferred into the proto-archaeal host's proteome.

Our results may indicate an underlying reason for this selectivity. Irrespective of the particular pathways and species-specific enzymes that are used, large-scale metabolic organization is essentially identical in all contemporary species, all of which possess a robust and error-tolerant scale-free network architecture[11]. In addition, other approaches for analyzing the functional capabilities of metabolic networks indicate that the complete metabolic network is under organizational constraints[22,23]. From this we can infer a selective pressure during eukaryote evolution that limited their incorporation of bacterial metabolic enzymes in order to maintain the existing favorable metabolic network architecture inherited from the proto-archaean host.

Methods

Database preparation. For our analyses of metabolic and information transfer pathways, we used the "Intermediate Metabolism and Bioenergetics" and "Information Transfer" portions, respectively, of the WIT database (http://igweb.integratedgenomics.com/IGwit/)[10]. This database predicts the existence of a biochemical pathway based primarily on the annotated genome of the organism combined with firmly established data from the biochemical literature. As of June 2000, this database provided description for 6 archaea (*Aeropyrum pernix, Archaeoglobus fulgidus, Methanobacterium thermoautotrophicum, Methanococcus jannaschii, Pyrococcus furiosus* and *Pyrococcus horikoshii*), 32 bacteria (*Aquifex aeolicus, Chlamydia pneumoniae, Chlamydia trachomatis, Synechocystis* sp., *Porphyromonas gingivalis, Mycobacterium bovis, Mycobacterium leprae, Mycobacterium tuberculosis, 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* and *Deinococcus radiodurans*), and 5 eukaryotes (*Emericella nidulans, Saccharomyces cerevisiae, Caenorhabditis elegans, Oryza sativa* and *Arabidopsis thaliana*) (note the absence of metazoan eukaryotes, including *Homo sapiens*, from the list). Complete genome sequences were available for many of these organisms, and incomplete sequences were available for *P. furiosus, 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, E. nidulans, O. sativa* and *A. thaliana*). The downloaded data were rechecked manually, and synonyms and substrates without defined chemical identity were removed.

Construction of network matrices and data sets. Biochemical reactions described within a WIT pathway 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. For a given organism with N substrates, E enzymes and R intermediate complexes, the full stoichiometric interactions for metabolism and information transfer were compiled into an $(N+E+R) \times (N+E+R)$ matrix, generated separately for each of the 43 organisms[11]. Four data sets, METAB/ENZ, METAB/SUBS, INFO/ENZ and INFO/SUBS, were created with 834, 1267, 115 and 395 rows, respectively, all with 43 columns. In each matrix, score x_{ij} indicates the number of times enzyme or substrate i is present in the corresponding pathway of organism j.

Enzyme and substrate ranking. Enzymes and substrates present in the metabolic and information transfer pathways of all 43 organisms were ranked based on the number of links each had in each organism, with incoming and outgoing links considered separately ($r = 1$ was assigned to the enzyme and substrate with the largest number of connections, $r = 2$ to the second most connected one, and so on).

Multivariate analyses. Neighbor joining (NJ), unweighted group average clustering (UPGMA), ordinal clustering (OC) and nonmetric multidimensional scaling (NMDS) were performed as previously described[12-15] (See Web Note B). All computations were made by the SYN-TAX 2000 software developed for WINDOWS systems[24].

Analysis of the effect of database errors. At the time of our analyses, the genomes of 25 of the 43 organisms had been completely sequenced (5 Archaea, 18 Bacteria, 2 Eukaryotes), and the remaining 18 were partially sequenced. Therefore, two major possible sources of error in the database could have affected our analysis: (a) erroneous annotation of enzymes and consequently of biochemical reactions, and (b) reactions and pathways missing from the database. For organisms with completely sequenced genomes, (a) is the likely source of errors; for those with incompletely sequenced genomes, both (a) and (b) are potential sources. We investigated the effect of database errors on the validity of our findings and found that our results are not affected (See Web Note C).

Note: Supplementary information is available on the Nature Genetics web site (http://genetics.nature.com/supplementary_info/).

Acknowledgments

We would like to acknowledge all members of the WIT project for making their invaluable database publicly available for the scientific community. Research at Eötvös University was supported by the Hungarian National Research Grant Foundation (OTKA), by the National Science Foundation at the University of Notre Dame, and at Northwestern University by grants from the National Cancer Institute.

Received 18 May; accepted 20 July 2001.

1. Darwin, C. *The Origin of Species* 6th edn (J. Murray, London, 1872).
2. Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Oxford University Press, Oxford, 1995).
3. Brooks, D.R. The nature of the organism. Life has a life of its own. *Ann. NY Acad. Sci.* **901**, 257–265 (2000).
4. Woese, C.R., Kandler, O. & Wheelis, M.L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**, 4576–4579 (1990).
5. Snel, B., Bork, P. & Huynen, M.A. Genome phylogeny based on gene content. *Nature Genet.* **21**, 108–110 (1999).
6. Rivera, M.C., Jain, R., Moore, J.E. & Lake, J.A. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. USA* **95**, 6239–6244 (1998).
7. Margulis, L. *Origin of Eukaryotic Cells* (Yale University Press, New Haven, CT, 1970).
8. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
9. Moreira, D. & Lopez-Garcia, P. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J. Mol. Evol.* **47**, 517–530 (1998).
10. Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**, 123–125 (2000).
11. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabási, A.L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
12. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
13. Sokal, R. & Sneath, P. *Numerical Taxonomy* (Freeman, San Francisco, 1973).
14. Podani, J. Explanatory variables in classifications and the detection of the optimum number of clusters. in *Data Science, Classification, and Related Methods* (ed. Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.H. & Baba, Y.) 125–132 (Springer, Tokyo, 1998).
15. Podani, J. *Introduction to the Exploration of Multivariate Biological Data* (Backhuys, Leiden, 2000).
16. Tourasse, N.J. & Gouy, M. Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. *Mol. Phylogenet. Evol.* **13**, 159–168. (1999).
17. Andersson, J.O. & Andersson, S.G. Insights into the evolutionary process of genome degradation. *Curr. Opin. Genet. Dev.* **9**, 664–671 (1999).
18. Penny, D. & Poole, A. The nature of the last universal common ancestor. *Curr. Opin. Genet. Dev.* **9**, 672–677 (1999).
19. Miyata, T. *et al.* Evolution of archaebacteria: phylogenetic relationships among archaebacteria, eubacteria, and eukaryotes. in *Evolution of Life: Fossils, Molecules, and Culture* (ed. Owasa, S. & Honjo, T.) 337–351 (Springer, Tokyo, 1991).
20. Doolittle, W.F. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307–311 (1998).
21. Lawrence, J.G. & Roth, J.R. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843–1860 (1996).
22. Schilling, C.H., Letscher, D. & Palsson, B.O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248 (2000).
23. Schuster, S., Fell, D.A. & Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.* **18**, 326–332 (2000).
24. Podani, J. *SYN-TAX 2000. Computer Programs for Data Analysis in Ecology and Systematics* (Scientia Publishing, Budapest, 2001).