

The three-domain classification of life forms into bacteria (B), archaea (A) and eukaryotes (E) is thought to be a major accomplishment of the evolutionary biology at the end of the 20th century. This classification, and the branching order (B(AE)), with the root of the tree between B and AE (1-4), are accepted throughout my work, but I also consider one qualification and one dissenting opinion now.

The qualification. The three-domain classification has been developed prior to complete genome sequencing, mostly on the basis of phylogenies of rRNA and of proteins involved in DNA replication, transcription, and mRNA translation. Analysis of complete genomes indicates that there exist two classes of genes. One set, dubbed “informational genes”, because it is dominated by the aforementioned genes involved in genome replication and expression, indeed tend to show the (B(AE)) topology. The other set, “operational genes” mostly involved in the intermediate metabolism, tend to follow ((BA)E) topology (5, 6). Note that eukaryotes also possess their own, relatively lately added, complement of bacteria-like genes, which are thought to have been acquired from bacterial precursors of mitochondria and chloroplasts and are mostly involved in assembly and function of these organelles, even as these genes currently reside in the nuclear genome. Yet, the presence of these genes does not sway the trees of most operational genes, and the ((BA)E) topology predominates. The conclusion from these observations is that, in of the early evolution of Life, at or near the time of emergence of the three domains, there were massive horizontal gene exchanges or genome mergers between the domains (5-9). Which exactly domains merged is the matter of debate (rooting problem). But in a computer-science sense, the Tree of Life may not be a tree at all, because it appears to contain the evidence of that merger close to the root

(reticulations that produces cycles, which are not compatible with the definition of a tree as a directed acyclic graph).

All this, however, has not much effect on the reconstruction of minimal and ancestral ribosome, because ribosomal proteins favor the (B(AE)) topology without any evidence of horizontal transfer between B and AE (excluding the relatively late acquisition of organellular genes, whose products are readily distinguished by their targeting peptides and by evolutionary affinities to specific bacterial groups from which the endosymbionts have originated). There are examples of likely horizontal transfer of ribosomal protein genes within the domain of Bacteria (*10, 11*), but they do not affect the conclusions of this work.

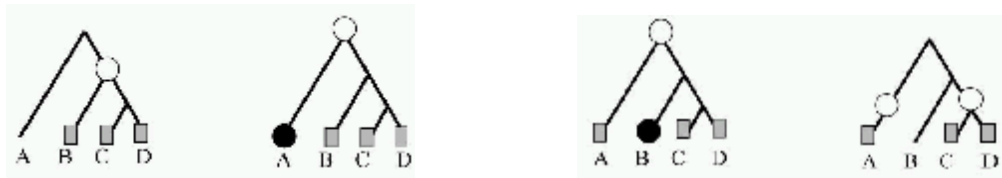
The dissenting opinion. A radically different relationship of three domains of Life has been proposed on the basis of a set of biochemical, paleontological, and cytological considerations (*12*), which seem to be corroborated by a cladistic analysis of specific shared insertions in selected protein-coding genes (*13*). Under this unorthodox scenario, archaea and eukaryotes are derived forms of bacteria. The radical difference of some ribosomal protein genes between AE and B, and the (B(AE)) topology support in a majority of them is interpreted as a result of accelerated evolution in AE (*12*).

This point of view has little effect on the reconstruction of Min1 and Min2 sets. Its impact on the Anc set is, however, large. All PPL1/2 and PPL3 proteins, unless they exhibit a very shallow phylogenetic tree indicative of overprinting, are now candidates for inclusion into the Anc set, as well as some of the PPL4 and PPL7 proteins with compatible phyletic patterns. This increases the number of proteins in the Anc set by 5-10

proteins, and, in the case of gene displacements, suggests a bacterial counterpart as the component of choice. I do not pursue this scenario further.

Phyletic patterns as the resource for reconstruction of the ancestral state. To quote Mirkin et al. (reference 14)

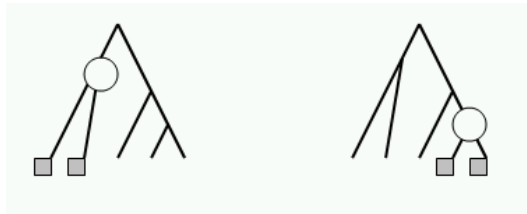
The reconstruction of the evolutionary scenario for an individual set of orthologous genes can be formulated as follows: given a species tree and a set of orthologs with a particular phyletic pattern (i.e. pattern of presence-absence of the species within the analyzed set of species; this set of species should be the same as in the tree), find the most parsimonious mapping of the set of orthologs on the tree. Such a mapping corresponds to the most parsimonious evolutionary scenario for the given set of orthologs, i.e. the scenario with the smallest possible number of events.



The count of events is illustrated in the two diagrams right above (adapted from Figures 2 and 3 in reference 14, full text can be found at <http://www.biomedcentral.com/1471-2148/3/2>). In the left-hand diagram, a gene is found in lineages B, C, and D, whose phylogeny is as shown. Such phyletic pattern may be explained either by a gene gain by the last common ancestor of B, C, and D (open circle; one event), or by gene appearance in the last common ancestor of all four species (open circle; first event) followed by gene loss in the lineage leading to A (black circle; second event). According to parsimony principle, the one-event explanation is better than the two-event scenario. In the right-hand diagram, a toss-up between two two-event scenarios (one gain + one loss vs. two gains) is shown. The details, including a rather

technical discussion of whether to count gene presence at the root as the event, and whether to score gains and losses differently, can be found in reference 14.

One parameter that the model of Mirkin et al. does not consider is the relative lengths of branches in different parts of the tree. Consider two cases of gene gain (single event in each case), mapped onto the same phylogenetic tree:



When deciding whether the gene has been present at the root, the parsimony principle will not distinguish between these cases – all it knows is that both distributions are explained by the same minimal number of events. If, however, the branch lengths are taken into account, the left-hand case is more suggestive of the ancestral presence of the gene, because it appears to persist in evolution since more ancient times (if only branch lengths can accurately reflect them).

We are developing a maximum-likelihood approach to ancestral gene content reconstruction that would utilize this information to improve the accuracy with which the ancestral gene content can be defined. Rates of gene gain and loss are estimated from the non-controversial portions of the phylogenetic tree, using modified Markov transition rate model for discrete morphological characters (15), and scenarios of gene presence-absence at the root are evaluated using likelihood ratio test (ARM and G.V.Glazko, manuscript in preparation). Quantitative details aside, there is just one ribosomal protein, L07E with PPL7 (Table 1 in the main text of the manuscript), which is not placed in the ancestral gene set under most of the weighted parsimony schemes of Mirkin et al. (14), but its

existence in LUCA is supported better than its absence in the likelihood ratio test.

Therefore, L07E joins the Anc set.

Other methods

Sequence homologs were detected with the PSI-BLAST program (16), with settings $-t=F$ (no correction for compositional bias), $-h=0.02$, $-F=F$, run until convergence. The most diverse members of each family were used as queries in additional rounds of search. Phyletic pattern information was from the Clusters of Orthologous Groups (COG) database at NCBI (17), supplemented with the results of the database search. Presence/absence of each gene in LUCA was evaluated either under parsimony assumption using the approach of Mirkin *et al.* (14), or using a maximum-likelihood model of gene gain and loss (see above). Phylogenetic trees for ribosomal proteins were constructed using the maximum-likelihood approach, as implemented in proml program in the Phylip package (18). For analysis and visualization of molecular structures, the programs SwissPDB Viewer (19) and Pymol (20) were used. Ribosome-specific databases RPG (21) and DRC (22) were used as additional sources of information on ribosomal proteins and their interaction with rRNA. Alignments of ribosomal RNAs from different species were constructed using the utilities at the Comparative RNA Web site (23).

References

1. Woese CR: Bacterial evolution. *Microbiol Rev* 1987, 51:221-271.

2. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, *et al.*: Evolution of the vacuolar H⁺-ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci U S A* 1989, 86:6661-6665.
3. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T: Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci U S A* 1989, 86:9355-9359.
4. Woese CR, Kandler O, Wheelis ML: Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A* 1990, 87:4576-4579.
5. Koonin EV, Mushegian AR, Galperin MY, Walker DR. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol.* 1997 Aug;25(4):619-37.
6. Rivera MC, Jain R, Moore JE, Lake JA. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A.* 1998 May 26;95(11):6239-44.
7. Woese CR. Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A.* 2000 Jul 18;97(15):8392-6.
8. Koonin EV. Horizontal gene transfer: the path to maturity. *Mol Microbiol.* 2003 Nov;50(3):725-7.
9. Rivera MC, Lake JA. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature.* 2004 Sep 9;431(7005):152-5.
10. Brochier C, Philippe H, Moreira D. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet.* 2000 Dec;16(12):529-33.

11. Makarova KS, Ponomarev VA, Koonin EV. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol.* 2001;2(9):RESEARCH 0033. Epub 2001 Aug 30.
12. Cavalier-Smith T. The neomuran origin of archaeobacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol.* 2002 Jan;52(Pt 1):7-76.
13. Gupta RS. The natural evolutionary relationships among prokaryotes. *Crit Rev Microbiol.* 2000;26(2):111-31.
14. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol.* 2003 Jan 06;3(1):2. Print 2003 Jan 6.
15. Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society (B)* 255, 37-45.
16. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
17. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* **4**, 41. Print 2003 Sep 11. (2003)
18. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. (2004).

19. Guex, N., & Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714-2723 (1997).
20. DeLano, W.L. The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA. <http://www.pymol.org> (2004).
21. Nakao, A., Yoshihama, M., & Kenmochi, N. RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res* **32**, Database issue D168-D170 (2004).
22. Baranov, P.V., Kubarenko, A.V., Gurvich, O.L., Shamolina, T.A., & Brimacombe R. The Database of Ribosomal Cross-links: an update. *Nucleic Acids Res* **27**, 184-185 (1999).
23. Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V., Muller, K.M., Pande, N., Shang, Z., Yu, N., & Gutell, R.R. The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and other RNAs. *BMC Bioinformatics*, **3**, 2 (2002).

Supplementary Figure: protein fold usage in minimal ribosome, % of all proteins

