Primer

The real 'domains' of life

David A. Walsh and W. Ford Doolittle

Six months ago, Alastair Simpson and Andrew Roger published in these pages a primer on "The real 'kingdoms' of eukaryotes". This Primer should be seen as a companion to theirs, addressing not only the currently accepted classification of prokaryotes, but also the inferred evolutionary relationships among prokaryotes — Bacteria and Archaea — and between them and eukaryotes. It may seem surprising in this postgenomic era that these are still areas of active research and vigorous controversy. The relationships are not simple ones, however, and there is legitimate disagreement, at the philosophical level, about how the complexities should be dealt with to produce the best 'natural classification'.

Deeply divided prokaryotes

The key molecular player in microbial classification has been the RNA component of the small subunit of ribosomes (SSU rRNA, or 16S/18S rRNA), which Carl Woese insightfully picked in the early 1970s as a convenient and reliable 'universal molecular chronometer'. His goal was nothing less than a global Tree of Life, relating all living things, but most immediately his purpose was to sort out the prokaryotes. In 1977, he and his postdoc George Fox were ready to announce to the world that these could be unequivocally divided into two very distinct groups, on the basis of SSU rRNA sequence. The first group comprised mostly well-studied organisms, such as *E coli*, cyanobacteria and anthrax, which they called 'eubacteria'. The second was made up of less well-known types, such as methanogens and (as they soon discovered) extreme halophiles and some thermophilic acidophiles, which they named collectively 'archaebacteria'.

That prokaryotes are diverse was no surprise, but that they could be so neatly divided into two, and only two, groups certainly was, and so not widely accepted until other characteristics that distinguished the domain *Archaea* from the domain *Bacteria* (as they are now called) were described. By the early 1980s, such traits were known to include: the possession of RNA polymerases more like their eukaryotic than their bacterial counterparts in subunit composition and sequence; some features of translation shared specifically with eukaryotes; insensitivity to most antibacterial antibiotics; and unique membrane glycerolipids composed of isoprenols ether-linked to glycerol-1-phosphate, those of bacteria and eukaryotes being fatty acids ester-linked to glycerol-3-phosphate. Ether-linked lipids have, however, now been found in several thermophilic bacteria, and fatty acids were recently detected in an archaeon, leaving only the stereoisomeric form of the glycerol phosphate backbone as a diagnostic tool to differentiate absolutely between archaeal and bacterial membranes.

In the early 1970s, only partial sequence information (catalogs of oligonucleotides generated by nucleases) could be obtained. Now, of course near complete genes are easily PCR-amplified, cloned and sequenced. The SSU rRNA database as of February 2005 included more than 125,000 entries! These continue to support the division of prokaryotes into two domains, each with subdivisions most commonly called 'phyla' (Figure 1). Archaea show so far only two or three major constituent groups (perhaps they should be 'kingdoms'): the Euryarchaeota, the Crenarchaeota and (possibly) the Korarchaeota. Among Bacteria there are at least 52 phyla; some of these turn out to correspond closely to divisions of bacteria recognized in pre-molecular sequence days by molecular and cellular phenotype alone, such as cyanobacteria and spirochaetes. Some unexpected groupings that could not be easily unified by phenotypic similarities include the *Chloroflexi* assemblage and the Proteobacteria subdivisions. Even for previously recognized phyla, SSU rRNA sequencing provides the advantage of quick identification and the ability to define within-phylum phylogenetic relationships down to the level of 'species' in a uniform way.

Furthermore, molecular sequencing does not require strain isolation and culturing, as phenotyping does. Culture-independent approaches, developed first in Norman Pace's lab, have revolutionized microbial ecology just as radically as Woese's vision and hard work transformed microbial classification. PCR amplification and sequencing of DNA prepared straight from environmental or clinical samples allows the identification of bacteria and archaea which have not been and possibly cannot be cultured — indeed which may have never been seen! Half the bacterial phyla are known only in this way, as is a basal group of Archaea, the Korarchaeota. Also, it was through sequencing of environmental DNA that we first learned that archaea are not all extremophiles: indeed, pelagic crenarchaeota make up 20% of the picoplankton in the world ocean.

Although there are some fairly well-supported groupings of bacterial phyla in the SSU rRNA tree, the tree overall shows a 'star phylogeny' for bacteria. It is as if most bacterial phyla emerged over a very short period of evolutionary time, a 'big bang' adaptive radiation (analogous to the Cambrian explosion of metazoan body design) made possible, perhaps, by refinements in efficiency and integration of the cellular machinery. But a serious alternative explanation is just that the tree is *unresolved*: there is too little phylogenetic signal in genes to allow reconstruction of such ancient evolutionary branchings.

The Bacterial/Archaeal dichotomy has been widely accepted for more than two decades, during which there has been steady progress in the molecular biology of representatives of both domains, most striking for the Archaea, about which little was previously

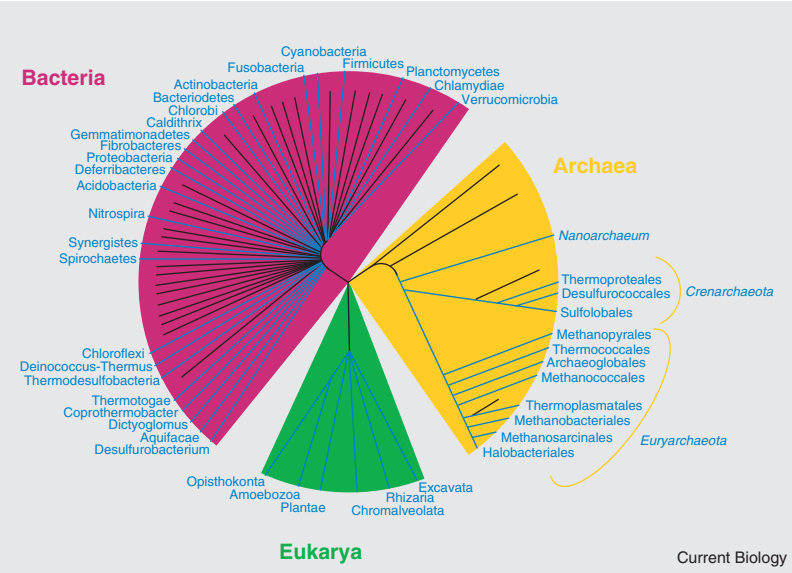

Figure 1. A diagrammatic representation of the organization of life into three domains based on SSU rRNA gene sequence similarity.

Blue branches represent those groups with cultured representatives; black branches represent groups only known from culture-independent environmental studies. The branching orders within the Archaea, Bacteria and Eukarya are based on those presented in Rappe and Giovannoni (2003), Forterre *et al.* (2002) and Simpson and Roger (2004), respectively.

known. What has this added to our understanding of their unique character?

Shared, unique and mosaic nature of prokaryotic informational systems

Features common to the molecular biology of bacteria and archaea are many, including: a typically (but not always) circular chromosome(s); absence of spliceosomal introns; organization of many genes into operons (sometimes with homologous genes in the same order); and, compared to most eukaryotes, simple cellular organization. It is the components of the information systems — transcription, translation and replication — that most surely distinguish the two prokaryotic domains, as often as not showing a strong affinity between archaea and eukaryotes.

For instance, archaeal RNA polymerases resemble a simplified version of eukaryotic RNA polymerase II, and archaea possess homologs of the associated eukaryotic transcription factors TBP (TBP), TFIIB (TFB), and TFIIEα (TFE). TBP and TFB are required for efficient promoter recognition and transcription initiation; TFE has a stimulatory role at some archaeal promoters, facilitating TBP binding to archaeal TATA box sequences. In some archaea, multiple copies of TBP and TFB may regulate gene expression, but other aspects of transcription regulation appear bacterial in character. A recent study of the phylogenetic distribution of transcription factor families in bacteria and archaea identified nine conserved in both, and several instances of transfer of transcription regulator genes between bacteria and archaea.

The core components of translation in archaea are most similar to those of eukaryotes. According to a recent survey, 33 ribosomal proteins are shared uniquely among Archaea and eukaryotes. In contrast, no ribosomal proteins are shared between Bacteria and eukaryotes that are not also found in Archaea, nor are there any that are shared between Bacteria and Archaea that are not also found in eukaryotes. Translation initiation in archaea and eukaryotes starts with methionine and not N-formyl-methionine as in bacteria. Bacterial features of archaeal gene expression include polycistronic uncapped mRNAs and translation initiation that requires base-pairing between a Shine-Dalgarno sequence at the 5′ end of the mRNA and a complementary sequence in the 16S rRNA. A second archaeal mechanism for translation initiation functions with leaderless mRNAs, more akin to the eukaryotic pathway.

Certain members of the Euryarchaeota possess histones, which they use to compact DNA. The evolutionary homology of these proteins to eukaryotic proteins of the same name seems beyond dispute, although they are shorter, corresponding to the core nuclear histone fold and lacking tail extensions. Studies in *Methanococcus fervidus* showed that archaeal nucleosomes assemble into a tetramer analogous to eukaryotic $[H3\text{-}H4]_2$ tetrasomes. Interestingly, in *Thermoplasmatales*, histones are replaced by small basic chromatin proteins (HU) found in bacteria. Crenarchaeaotes appear to have species-specific solutions to chromatin compaction, such as the Sul7d family of DNA binding proteins in *Sulfolobus*, and Alba, an abundant DNA-binding protein characterized in *Sulfolobus* and found in other thermophilic archaea and some eukaryotes. Alba coats double-stranded DNA without significant compaction, and it represses transcription. Lysine acetylation lowers Alba's affinity for DNA; *Sulfolobus* has a homolog of the eukaryotic histone deacetylase Sir2, which specifically deacetylates Alba, analogous to eukaryotic modulation of transcription by covalent modification of chromatin proteins.

Bacterial replication occurs from a single origin of replication, whereas eukaryotic genomes have multiple origins of replication. Archaea have a similar genome architecture to bacteria, usually consisting of a single circular chromosome, but — at least in *Sulfolobus* — with several replication origins. In bacteria, a family C DNA polymerase (DNA polymerase III) is the major replicative enzyme, while the replicative polymerases in eukaryotes are family B DNA polymerases (α, δ and ε).

Table 1. Identification of prokaryotic domain signature genes.

No. of orthologs present in	Archaeal genomic signature			Bacterial genomic signature		
	0% bacteria	≤10% bacteria	≤20% bacteria	0% archaea	≤10% archaea	≤20% archaea
100% archaea	28.1, sd = 4.1	43.2, sd = 4.2	48.3, sd = 4.8	–	–	–
≥90% archaea	48.0, sd = 4.6	77.4, sd = 7.6	89.2, sd = 10	–	–	–
≥80% archaea	64.5, sd = 6.1	105.7, sd = 11.13	123, sd = 14.6	–	–	–
100% bacteria	–	–	–	13.6, sd = 1.7	15.7, sd = 1.8	17.1, sd = 1.8
≥90% bacteria	–	–	–	32.3, sd = 2.4	40.7, sd = 3.6	43.2, sd = 3.5
≥80% bacteria	–	–	–	47.8, sd = 5.3	64.9, sd = 7.4	70.5, sd = 7.9

Genomic signatures were identified using the Group-specific genome query available at www.neurogadgets.com/bioinformatics.php with an inclusion threshold of $1.0e^{-10}$ and an exclusion threshold of $1.0e^{-5}$. Reported are means and standard deviations (sd) from the perspective of different query genomes. A total of 21 archaeal genomes and 195 bacterial genomes were included in the analyses.

Consistent with their other eukaryotic tendencies, archaea contain one to three family B DNA polymerases: their role as replicative polymerases is still to be confirmed. Intriguingly, a novel family D DNA polymerase has been identified in the euryarchaeote *Pyrococcus furiosis*; this enzyme prefers RNA-primed instead of gapped DNA as a polymerization substrate and exhibits higher processivity than the family B polymerase from the same organism, suggesting it may be the major replicative enzyme in this euryarchaeote.

Issues of diversity and unity

Woese has argued that lumping Bacteria and Archaea together as prokaryotes disrespects their fundamental differences. There is similarly much structural and functional diversity within each domain, and we have already noted several exceptions to generalizations about the eukaryotic character of archaeal information systems. We can expect to discover more as we learn more. Two bacterial surprises in just the last few years involve the Planctomycetes and the Verrucomicrobia. Members of the former have (like Archaea) cell walls lacking peptidoglycan and (like eukaryotes) membrane-bounded nucleoids. Members of the latter contain genes for the eukaryotic cytoskeletal protein tubulin and homologs of a few additional genes otherwise considered restricted to eukaryotes. An archaeal surprise was the discovery of a new archaeal phylum, so far represented by a single species, *Nanoarchaeum equitans,* a parasite of the hyperthermophilic crenarchaeote *Ignicoccus*. The genome of *N. equitans* is highly reduced (<500 kb), lacking genes for lipid, cofactor, amino acid and nucleotide biosynthesis. Phylogenetic analyses based on concatenated ribosomal proteins place *N. equitans* at the base of the Archaea. Whether it represents a novel 'primitive' archaeal phylum or is a highly derived euryarchaeote or crenarchaeote is still a matter of debate.

Observationally based information on structural and functional diversity within and between Bacteria and Archaea is still sparse and hard to obtain for organisms not seen as vital to health or wealth. So completed genome sequences may offer the best and least biased measure of the dissimilarity of the two prokaryotic domains, and the similarity of the organisms within each. The current availability of >200 bacterial and >20 archaeal genome sequences makes it possible to assess functional diversity within and between domains in a more systematic way. We can ask, for instance, how many genes have orthologous copies in all archaea and no bacteria (or vice versa), to define domain-specific 'genomic signatures'. Or, if we want to discount the effects of gene loss in obligate parasites, such as *Nanoarchaeum*, and the products of between-domain lateral gene transfer (LGT), we can ask how many genes have orthologs in at least 90% of archaea and no more than 10% of bacteria, and so forth. The results of such an exercise are shown in Table 1. There *are* domain-specific signatures, although they include relatively few genes (a few percent of the number of genes in most genomes). And it is overwhelmingly *informational* genes which differentiate domains, either by being present in only one, or by showing, in phylogenetic reconstructions, the deep Bacterial–Archaeal division. Indeed the most believably robust universal phylogenies are those based not on single genes (which may have too little phylogenetic signal) but multi-gene concatenated sequences of ribosomal and other proteins of the translational and transcriptional machinery.

On the other hand, many *operational* genes (for anabolism and catabolism, structure and communication) are patchily distributed within both prokaryotic domains, in a fashion that can best be explained by inter-domain LGT. One-quarter of the genes of the hyperthermophilic bacterium *Thermotoga maritima* appear to be derived from (have their closest match in) archaea, while nearly a third of the genes of the euryarchaeote *Methanosarcina mazei* look to be bacterial. There is unquestionably a diverse pool of genes functioning in energy metabolism, the formation and degradation of small metabolites, regulation of gene expression and such key environmental processes as nitrogen fixation, from which both bacteria and archaea have drawn. We might see the genes in this pool as software, readable by two different kinds of hardware — the bacterial and archaeal

information systems — and 'belonging' to neither prokaryotic domain.

Rooting the tree, the three domain view of life and the prokaryote/eukaryote dichotomy

The third domain, of course, is Eukarya, the subject of Simpson and Roger's recent Primer. What is its relationship to the other two, in evolutionary terms? The consensus view, found in most textbooks, is that the earliest evolutionary branching separated Bacteria from the lineage which was later to produce Archaea and Eukarya. This consensus is supported by phylogenetic trees based on certain anciently duplicated proteins (present in multiple copies in the common ancestor) and on the common-sense notion that the differences between archaeal/eukaryotic and bacterial information systems reflect their separate evolution from a primitive ancestral state. But several serious scientists have suggested that the archaeal/eukaryotic informational machinery is in fact the ancestral state, of which bacteria have a secondarily simplified form. Others argue that the deepest branching within the universal tree lies within the known bacteria, Archaea and eukaryotes having sprung from within the Gram-positive bacteria.

We adopt the consensus, but do not consider it proven, and note that basing an understanding of the relationship between Life's three domains on a small, albeit important, subset of their genes (those of transcription, translation, and replication) disenfranchises the majority, for which inter-domain sharing may be the rule. Martin and collaborators, for instance, recently noted that "approximately 75% of yeast genes having homologues among the present prokaryotic sample share greater amino acid sequence identity to eubacterial than to archaebacterial homologues".

Woese and others have argued repeatedly that the prokaryote/eukaryote dichotomy, which still dominates much of the biological literature and the thinking of many biologists, should be expunged from both, as it underplays the important differences between Bacteria and Archaea, and ignores cladistic principles (by uniting two groups, Bacteria and Archaea, which are on different sides of the universal tree's deepest branching, while excluding Eukarya which are, according to that tree, Archaea's closest relatives). One could say in the prokaryote/eukaryote dichotomy's defense that it was never intended to have cladistic meaning, but was rather a description of two types of cellular organization, and it remains valid as that. Eukaryotic cells, except a few highly derived parasitic forms, boast a degree of internal complexity (cytoskeleton, endomembrane systems, membrane-enclosed energy generating organelles) not so far found in either Archaea or Bacteria.

Indeed, Simpson and Roger point out that the last common ancestor of all surviving eukaryotes was likely such a complex cell. All current theories about the origin of eukaryotes see them as chimeras arising from the coming together of different prokaryotic cellular lineages, both archaeal and bacterial, *via* some intermediate symbiotic association. So eukaryotes differ from prokaryotes not only in their complexity but in the evolutionary process which gave rise to them, and because of that process, are outside traditional cladistic treatment — there is no *single* older prokaryotic lineage to which they can be assigned. Prokaryotes are likely overall much less uniform in their basic cellular and molecular properties but, because of the existence among them of two (largely) distinct types of information system 'hardware', are sensibly divided in two. Still, as cells there would never be difficulty in recognizing either type as not eukaryotic. Indeed, Woese has suggested that the term 'prokaryote' should be replaced by 'non-eukaryote'. We suspect that this is what, for most biologists, 'prokaryote' has always meant.

Acknowledgments

We thank Robert L. Charlebois for his help in identifying domain signature genes and Thorsten Allers for critical reading of the manuscript. We also acknowledge Genome Atlantic and The Canadian Institute for Health Research for financial support.

Further reading

Allers, T., and Mevarech, M. (2005). Archaeal genetics - the third way. Nat. Rev. Genet. *6*, 58–73.

Charlebois, R.L., and Doolittle, W.F. (2004). Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res. *14*, 2469–2477.

DeLong, E.F., and Pace, N.R. (2001). Environmental diversity of bacteria and archaea. Syst. Biol. *50*, 470–478.

Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., *et al.* (2004). A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol. Biol. Evol. *21*, 1643–1660.

Forterre, P., Brochier, C., and Philippe, H. (2002). Evolution of the Archaea. Theor. Popul. Biol. *61*, 409–422.

Perez-Rueda, E., Collado-Vides, J., and Segovia, L. (2004). Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. Comput. Biol. Chem. *28*, 341–350.

Rappe, M.S., and Giovannoni, S.J. (2003). The uncultured microbial majority. Annu. Rev. Microbiol. *57*, 369–394.

Simpson, A.G., and Roger, A.J. (2004). The real 'kingdoms' of eukaryotes. Curr. Biol. *14*, R693–R696.

Staley, J.T., Bouzek, H., and Jenkins, C. (2005). Eukaryotic signature proteins of *Prosthecobacter dejongeii* and *Gemmata* sp. Wa-1 as revealed by in silico analysis. FEMS Microbiol. Lett. *243*, 9–14.

Waters, E., Hohn, M.J., Ahel, I., Graham, D.E., Adams, M.D., Barnstead, M., Beeson, K.Y., Bibbs, L., Bolanos, R., Keller, M., et al. (2003). The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proc. Natl. Acad. Sci. USA *100*, 12984–12988.

Woese, C.R. (2004). A new biology for a new century. Microbiol. Mol. Biol. Rev. *68*, 173–186.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA *74*, 5088–5090.

Genome Atlantic, Department of Biochemistry and Molecular Biology, Dalhousie University, 5850 College St., Halifax, Nova Scotia, B3H 1X5 Canada.
E-mail: dawalsh@dal.ca and ford@dal.ca