# Microbial evolution: The new synthesis

W.F. Doolittle

Program in Evolutionary Biology, Canadian Institute for Advanced Research, Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia

## ABSTRACT

Over the last two decades, molecular phylogeneticists have built up an understanding of relationships within and between prokaryotic and eukaryotic microbial organisms which seems to have explanatory and predictive power. There is at the moment general confidence in the distinctness and coherence of the Archaea and Bacteria, in the notion that eukaryotic cells arose by an increase in internal complexity from an archaea-like ancestor, in the (a-proteobacterial) endosymbiotic origin of mitochondria and in the primitive status of certain deeply diverging lineages of protists. Come the millennium, however, we are likely to see our confidence shaken, as more genome sequences appear and are compared. Archaeal and eukaryotic genomes are unexpectedly chimeric. It is not clear whether there are any genes whose products are so tightly integrated in cellular function that they cannot be replaced. The concept of genomic lineages may be inappropriate for understanding cellular evolution, but it is not obvious what will replace it.

## Introduction

The goal of this Opening Plenary is to present an overview of current opinion about the first branches of the tree of Life, in other words about the early evolution of prokaryotes some 3.5 billion years ago and the derivation from them of eukaryotes, perhaps a billion years later. It has four sections. The first, corresponding to the first 60 years of this century, describes the futility of attempts to construct a reasonable prokaryotic phylogeny from comparative morphology and physiology, and the consequent appeal of molecular sequence methods. The second, summarizing most of the last three decades, documents the success of such methods, especially those based on ribosomal RNA, and the forging of a consensus about the structure of the universal tree. The third section, covering just the last few years, recounts how new gene and genome sequences challenge this consensus, in particular providing many examples of noncongruent trees, which could mean that lateral gene transfer has been more important and more frequent than we had recently imagined. The final section looks to the future, and asks whether frequent lateral transfer in fact fatally compromises the entire deep phylogenetic enterprise.

## Morphology and physiology fail

Pre-molecular microbial taxonomists sought to group prokaryotes and unicellular eukaryotes into natural kinds based on morphology detectable in the light microscope or biochemical parameters of growth. Pre-molecular microbial phylogeneticists (often the same people) sought to reconstruct the genealogical relationships between such groups. No consensus emerged and by the mid 1940s, Stanier and van Niel, long boosters of cell morphology, gave up the attempt - the latter writing that "morphological criteria have, on

***Microbial Biosystems: New Frontiers***
*Proceedings of the 8th International Symposium on Microbial Ecology*
*Bell CR, Brylinsky M, Johnson-Green P (ed)*
*Atlantic Canada Society for Microbial Ecology, Halifax, Canada, 1999.*

the whole, a far more restricted significance [for microbes than for animals and plants] and do not justify any conclusions about phylogeny" [1]. Similarly phylogenetic efforts based on metabolism, which can be remarkably uniform within groups, foundered for lack of any way to choose between competing schemes for how one metabolism evolves from another. Confounding both was a regrettable tendency (still widespread) to regard some living organisms as being in some essential way more "primitive" than others retaining, in all or most of their features, an especially strong resemblance to ancestral forms.

Although finally apostate on microbial genealogy, Stanier and van Niel thought that one very clear distinction could be drawn within the microbial world -- between all prokaryotes and all eukaryotes. This dichotomization, first proposed by Chatton in 1937, required supporting data from electron microscopy and molecular biology because Stanier and van Niel realized, "the differences between eukaryotic and prokaryotic cells are not expressed in any gross features of cellular function: they reside rather in differences with respect to the detailed organization of the cellular machinery" [2].

These differences mostly corresponded to complex structures eukaryotes have and prokaryotes do not. In 1970, Stanier [3] proposed that they arose in consequence of satisfying a particular selective pressure; the pressure to engulf entire bacterial cells as food, through endocytosis (phagocytosis). By evolving a cytoskeleton and internal compartmentalizing membranes, early "protoeukaryotes" finessed the need to evolve a range of energy metabolisms: they could just eat the results. And sometimes, if digestion was delayed, food bacteria might become endosymbionts. Thus Stanier's vision of the protoeukaryote as a newly arisen phagocyte, coupled with Lynn Margulis' assertion that mitochondria and plastids are the degenerate descendants of bacterial endosymbionts, became the modern endosymbiont hypothesis for the origin of eukaryotes [4].

## Molecular methods and the hegemony of rRNA

Since these notions derived from ultrastructural comparisons, their best confirmation had to come from independent data, in particular molecular sequences. Zuckerkandl and Pauling, in 1965, argued that sequence data were to be preferred to organismal phenotypes, because the order of bases in DNA (or of amino acids in proteins) is deeper (in some ontological sense) than phenotype, and differences in sequence are more objectively quantitated and converted into evolutionary distances than are phenotypic resemblances [5].

In the 1960s, all we could sequence were proteins. The most extensive deep phylogenies were based on cytochromes and ferredoxins, markers of the evolutionary histories of mitochondria or plastids and the respiring bacteria or cyanobacteria thought to have given rise to them. Indeed, such relationships were well supported by these proteins and some general features of bacterial phylogeny could be discerned. However, precisely because they tracked organellar history (with genes transferred early from endosymbiont to nuclear genomes), cytochromes and ferredoxins could tell us nothing about the origin of the first endosymbiont's host: the protoeukaryote.

Other proteins, for example translation elongation factors which might have helped, had not been sequenced. So in their bold 1978 attempt to produce a universal tree, Schwartz and Dayhoff used 5S ribosomal RNA (rRNA) to link prokaryotes and eukaryotes [6]. This molecule ultimately proved too small and conservative for reliable phylogenies, but 16S rRNA, championed by Woese and his collaborators, has since conquered the field. This molecule (or the cognate gene) has been sequenced from many thousands of species and

16S phylogeny provides the scaffold on which evolutionary theories of all sorts are hung. It also provided the first step in the identification of the lineage of the protoeukaryotic host.

16S rRNA showed that prokaryotes are deeply divided, into Bacteria and Archaea - so deeply in fact that Woese argued that the prokaryote/eukaryote dichotomy should be de-emphasized, the Bacteria/Archaea/Eukarya trichotomy becoming the first organizing principle of biology. Mayr takes the contrary position. This debate may never end because it is not about the facts but about which of the facts are most important [7]. In any case, once we knew there were two deeply divided groups of prokaryotes, we could ask, "to which group did that brave prokaryote belong which first set off down the path to acquire phagocytosis and thus become a protoeukaryote?" This is the same as asking "where is the root of the universal tree?", or "which of Life's three domains diverged first?". Ribosomal RNAs could not give us a universal root, because they provide no outgroup. For this we need sequences of a universally distributed gene which is the product of a gene duplication which occurred prior to the last common ancestor of all Life, so that sequences of its duplicate (its paralog) can be used as outgroups. In 1989, Miyata's and Gogarten's groups promoted elongation factor and ATPase genes as such genes. We now have several others, almost all coding for components of the transcription or translation apparatus. Mostly, these genes support a rooting between Bacteria on the one hand and Archaea and eukaryotes on the other [8]. In other words, the protoeukaryotic host's closest prokaryotic relatives are archaea.

Since most of the criteria by which we traditionally distinguish eukaryotes from prokaryotes derive from comparisons between eukaryotes and bacteria, we should expect that archaea will exhibit some "eukaryotic features". Indeed they do: archaeal replication, transcription and translation machinery resemble those of eukaryotes not only "quantitatively" (in the amino acid sequences of their components) but "qualitatively" (in the numbers and detailed functions of those components). Archaeal genome sequences reveal homologs of many eukaryotic transcription factors and replication proteins never found in bacteria, and even before genome sequencing, we knew of archaea/eukaryote-specific ribosomal proteins. However, archaeal genome sequences do not show genes for the components of the features of cellular ultrastructure; the cytoskeleton and endomembrane system - by which Stanier and van Niel first sought to distinguish prokaryotes from eukaryotes. In what organisms might we expect to find these still in a nascent state?

For much of the last decade, we thought we might find primitive stages in eukaryotic cellular evolution in the "Archezoa", (presumed) primitively amitochondriate protists. Several authors (most notably Cavalier-Smith) had suggested that some descendants of the protoeukaryote host which had never harbored the endosymbiotic bacteria that were to become mitochondria might still be with us [9]. They might, for instance, be the mitochondria-free and relatively simple unicells we call metamonads (like Giardia), parabasalians (like Trichomonas), and microsporidia (for instance Encephalitozoon). When these "archezoal" lineages were shown, by SSU rRNA reconstructions, to comprise the deepest eukaryotic branches of the universal tree, this notion seemed to be confirmed. If our beliefs about the archaea-like character of the first protoeukaryote host are correct, archezoal genomes should be made up only of genes taken from the common archaeal/eukaryotic branch, suitably modified by selection for eukaryotic-like cellular functions.

## Our current discontents

Now (in just the last year or two) this "Archezoa hypothesis" seems disconfirmed, and doubly so. One of the three supposedly deep groups, the microsporidia, looks to have been placed so low on the tree because of "long-branch attraction" and other treeing artifacts. Most likely microsporidia are degenerate fungi. The deep branchings of the other two are also in doubt, although we don't know where else to put them. Furthermore all three lineages, and as far as we know all existing eukaryotes, have been tarred with the mitochondrial brush - all bear at least one and almost certainly more nuclear genes of unquestioned mitochondrial/a-proteobacterial origin [10]. In addition, there appear to be bacterial genes in all eukaryotic genomes many serving decidedly non-mitochondrial functions. No one now doubts that there was an early contribution by bacteria to the nuclear genomes of the earliest eukaryotes. What we do not know is how extensive this contribution is or how it was made.

As if that were not bad enough, archaeal genomes also look to be heavily "contaminated" with bacterial genes. There are several gene phylogenies which show archaea mixed in with bacteria (often Gram positive bacteria). Moreover, comparisons of the available complete archeal genome sequences reveal many "bacterial genes". That is to say genes that are more similar to their bacterial than their eukaryotic homologs or that have no eukaryotic homologs or that are known only from one archaeon and several bacteria. Archaeoglobus is a particularly good genome for examples [11]. Not surprisingly, as a chemoheterotroph it bears many genes for import and utilization of exogenous substrates that the autotrophic archaea, whose genomes were first sequenced, lack. Surprisingly many of these genes are likely to be traceable to specific sources within the bacteria, and thus to have been acquired by "lateral gene transfer". An increasingly popular view is that Archaea can be defined as a monophyletic assemblage by genes for transcription and translation (the same genes which show them to be sisters to eukaryotes). Genes for metabolic functions are more likely to be shared with bacteria, part of an exchangeable pool of useful biochemical information [12].

Older readers will recall that the notion of a shared prokaryotic gene pool was also popular in the late 1960s and 1970s when antibiotic resistance transfer factors and plasmids in general first came under intense scrutiny. It fell out of favor in the 1980s because in many instances lateral gene transfer had been invoked to rationalize poorly resolved gene trees, and (perhaps) because the molecular phylogenetic enterprise is futile if the genes do not behave. Although it is still too hard to separate signal from noise in genome by genome comparisons, and too easy to attract attention by making radically deconstructive claims, we do need to worry about what lateral gene transfer could mean for phylogeny (and taxonomy).

## What could extensive lateral gene transfer mean to us?

We could regard lateral gene transfer as simply a nuisance: exchange of certain genes between lineages means that these genes can not be trusted to give us the true phylogeny of organismal lineages. But there still is a true phylogeny reconstructable from some gene sequences. But what genes can we trust? Most would favor those of the transcription and translation apparatus for both bad reasons and good. Two bad reasons are that (i) the current universal phylogenies and our whole metataxonomic vocabulary are already based on such genes, and (ii) most of us were trained as molecular biologists and thus regard

such molecules as more important (more biologically fundamental) than enzymes of intermediary metabolism, say. Two better reasons would be that (i) transcription/translation genes are, because their products interact so extensively with each other and other macromolecules, unlikely to be successfully integrated individually into new genomic environments and (ii) molecular phylogenetic experience is that more gene phylogenies agree (and agree more unambiguously) with rRNA trees than with trees based on other molecules.

The first of these "better" reasons needs further examination. There are other multiprotein complexes in the cell (respiratory and photosynthetic complexes for instance) which are just as highly integrated and interactive. Transcription and translation genes are often clustered and could be transferred together. Antibiotics targeted against transcription/translation components could provide strong selection for transfer of resistant versions even if their initial integration is inefficient. And anyway, there are in fact well documented phylogenetic disagreements between rRNAs and RNA polymerases. The second, empirical, better reason may be the best. Certainly many gene phylogenies do agree very extensively with those based on 16S rRNA: *rec*A is a good example [13]. In the long run, when we have universal phylogenies derived from sequences of hundreds of different genes, we should be able identify some which always (or almost always) agree. But we should not play favorites in such analyses: at the moment rRNA is the gold standard and disagreements are often rationalized away.

No matter how such phylogenetic meta-analyses turn out, we may need to rethink the relationship between lineages of genes, lineages of organisms and prokaryotic taxonomy. Systematics is an older and different practice than phylogeny. The relatively recent notion that prokaryotic taxonomy should rest on (in fact be replaced by) a single phylogeny derived from gene sequences depends on a correspondence between gene lineages and organismal lineages which does not hold for all genes and can therefore only be arbitrarily assumed to hold for any particular gene. Organismal lineages defined by only one or a few of their genes are not the sort of individual-like entities most molecular phylogeneticists had in mind when they set out to reconstruct the universal tree of Life. Such lineages are more like athletic teams or symphony orchestras, constant in name but variable in essential constitution. We can probably live with this but need to recognize the provisional and ultimately limited meaning of the words we use to describe microbial taxa at the highest levels.

## References

1. van Niel CB (1946). The classification and natural relationships of bacteria. Cold Spring Harbor Symp Quant Biol 11: 285-300.
2. Stanier RY, van Niel CB (1962). The concept of a bacterium. Arch Mikrobiol 42:17-35.
3. Stanier RY (1970). Some aspects of the biology of cells and their possible evolutionary significance. Symp Soc Gen Microbiol 20:1-38.
4. Doolittle WF (1998). A paradigm gets shifty. Nature 392:15-16.
5. Zuckerkandl E, Pauling L (1965). Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel, eds. Evolving Genes and Proteins. Academic Press, New York, pp. 97-166.
6. Schwartz RM, Dayhoff MO (1978). Origins of prokaryotes, eukaryotes, mitochondria and chloroplasts. Science 199:395-403.

7. Woese CR (1998). Default taxonomy: Ernst Mayrs view of the microbial world. Proc Natl Acad Sci USA 95:11043-11046.
8. Brown JR, Doolittle WF (1997). Archaea and the prokaryote-to-eukaryote transition. Microbiol Mol Biol Rev 61:456-502.
9. Cavalier-Smith T (1983). A 6-kingdom classification and a unified phylogeny. In Endocytobiology II. Schwemmler, W. and Schenk, H. E.A. eds. pp. 1027-34. De Gruyter, Berlin.
10. Roger AJ, Clark CG, Doolittle WF (1998). A possible mitochondrial gene in the early branching amitochondriate protist *Trichomonas vaginalis*. Proc Natl Acad Sci USA 93:14618-14622.
11. Doolittle WF, Logsdon JM Jr. (1998). Archaeal genomics:do archaea have a mixed heritage? Curr Biol 8:R2009-R211.
12. Woese CR (1998). The universal ancestor. Proc Natl Acad Sci USA 95: 6854-6859.
13. Eisen J (1995). The *rec*A protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of *rec*As and 16S rRNAs from the same species. J Mol Evol 41: 1105-1123.