Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 8797–8799, August 1996

Commentary

At the core of the Archaea

W. Ford Doolittle

Canadian Institute for Advanced Research and Department of Biochemistry, Dalhousie University, Halifax, NS Canada B3H 4H7

There are three kinds of taxonomists, according to Ernst Mayr (1). *Pheneticists* group species by overall phenotypic similarity, renouncing evolutionary theory and explanation. *Cladists*, conversely, concern themselves exclusively with genealogy. Phenotypic resemblance between two taxa (lizards and crocodiles as reptiles, for instance), counts for nothing if one shares a more recent ancestor with a quite different sort of creature, as crocodiles do with birds. *Evolutionary taxonomists*, amongst whom Mayr includes himself and Charles Darwin, are compromisers, taking into account both branching order and "degree of difference"—phenotypic coherence makes reptiles real for them even though one branch of the reptilian tree bears birds.

Molecular phylogeneticists *ought* to be cladists, if not in method, at least in philosophy. They do sometimes infer relationships from measures of molecular similarity, but true genealogical trees are what they are after, and in any case there is little in the molecular sequences they deal with which speaks directly to the sort of organismal features of interest to pheneticists or evolutionary taxonomists.

However, for most of the 30 years since Zuckerkandl and Pauling (2) first suggested that sequences of molecules could be used to reconstruct evolutionary history, molecular phylogeny has been largely the handmaiden of evolutionary taxonomy, in Mayr's sense. For instance, in the 1970s, globin sequences were used to confirm and extend inferences about the tempo and mode of vertebrate evolution previously drawn from paleontology and comparative anatomy, while ferredoxin and cytochromes helped impose some rough taxonomic order on the bacteria, and underpinned tests of the endosymbiont hypothesis for the origin of eukaryotic cells (3). At issue in such exercises were the origins of and relationships between recognized groups about which we already had theories: molecular phylogeny was just the tool with which we tested them.

Molecular phylogeny began to move over into the driver's seat when a single molecule, small-subunit ribosomal RNA (16S, 18S, or SSU rRNA) won acceptance as the favored "molecular chronometer." Its hegemony is due in large part to the strong case, made both in argument and evidence by Carl Woese (4), that this was the all-around best choice. The *arguments* were that SSU rRNA is (*i*) universal, since all prokaryotic, eukaryotic nuclear, plastid, and mitochondrial genomes encode it; (*ii*) profoundly conservative in function and rate of change; and (*iii*) unlikely to be exchanged between lineages by "horizontal gene transfer," because its function is so fundamental and so dependent on so many intermolecular interactions. The *evidence* was that (*i*) when we knew what to expect, rRNA usually delivered the phylogenetic goods, and (*ii*) when it surprised us with unexpected relationships, subsequent work (other sequences, cell biology, and biochemistry) usually endorsed its conclusions.

The biggest of the surprises, of course, was the existence of the Archaea ("archaebacteria"). A collection of already-known but little-studied, difficult-to-classify, and superficially quite different prokaryotes—methanogens, halophiles, and extreme thermophiles—not only appeared to belong together, genealogically (4), but comprised an outgroup to all other prokaryotes (Bacteria, or "eubacteria"). Cladistics gave us this result, but its adoption in most textbooks and by most biologists hinged upon the enumeration of phenetic similarities confirming the "coherence" of the Archaea, such shared traits as isopranyl glycerol ether lipids (not fatty acid glycerol ester lipids), peptidoglycan-free cell walls, and certain eukaryotic-like transcriptional and translational features (complex RNA polymerases, unformylated methionyl tRNA, resistance to antibacterial antibiotics) looming large in this regard (5).

The phenetic coherence of the Archaea has also been a major defense against the principle challenge to Woese's tripartite universal taxonomy, which is James Lake's (6) repeated claim that *some* of the archaebacteria (thermophiles like *Sulfolobus* which he calls "eocytes") share a more recent common ancestor with eukaryotes than with the rest of the Archaea. Now most analyses of archaeal sequences *do indeed* show a deep split between the former, which Woese *et al.* (7) calls Crenarchaeotes and the latter (Euryarchaeotes, including thermophilic and mesophilic methanogens and sulfur metabolizers as well as halophiles). But until recently, only Rivera and Lake's (8) description of a single insertion in elongation factor (EF)-1α genes as a derived feature shared by crenarchaeotes and eukaryotes argued strongly for his eocyte notion. In terms of overall phenotypic similarity, Archaea appear a coherent "natural group." Although we have increasing evidence for eukaryote-like functional features in archaeal transcription and translation systems (9), as far as we know these features are found in all Archaea.

Whichever way this issue settles itself (of which more below), it remains one in which both sides have made heavy use of both phenetic and cladistic criteria. Although molecular phylogeny now often generates the hypotheses and organismal biology is the tool with which they are tested, sequence data and cell biological and biochemical features are still being played off against each other, in the best tradition of evolutionary taxonomic argumentation.

Trees Without Organisms

Norman Pace, in a series of bold investigations begun more than a decade ago (10), has used SSU rRNA to move us out of evolutionary taxonomy, beyond any such playing-off of organismal biology and molecular cladistics. With the aid of the polymerase chain reaction (PCR) and primers designed against conserved regions of eukaryal, archaebacterial, or eubacterial SSU rRNAs, Pace and his collaborators can amplify rRNA-encoding DNAs (rDNAs) directly from the environment. They can make phylogenetic trees for organisms that no one has in culture, for which there is no literature of cell biological and biochemical characterization, indeed which no one has ever seen! The method is now in widespread use, and widely hailed as *the* approach to documenting, understanding, and exploiting biological diversity. It also allows, and demands, practice of the purest sort of cladism—there is no supporting or conflicting biology, aside from that which can be inferred from the ecology of the site of isolation.

In the *Proceedings* in 1994, Barns *et al.* (11) described 17 new archaeal rDNAs from a single not very big hot spring ("Jim's Black Pool", or "Obsidian Pool") in Yellowstone National

Park. All were crenarchaeal, but most were only distantly related to previously known members of this assemblage. Since many sequence types were recovered only once, Barns and coworkers predicted that the full phylogenetic diversity and depth of this warm little pond had yet to be fully plumbed. Two sequences (pJP27 and pJP78) branched below all known Crenarchaeotes, and they ventured that these might even fall below the Crenarchaeal/Euryarchaeal split, when further sequences could be added to improve the resolution of deep branches.

In this issue of the *Proceedings*, Barns *et al.* (12) tell us that both predictions may be confirmed. Twenty-one new sequences have been obtained in a second visit to the same pool—2 euryarchaeal sequences resembling the marine thermophilic sulfate-reducing euryarchaeote *Archaeglobus fulgidis* and 19 representing new crenarchaeotes, most branching below named crenarchaeal species and many below isolates from the 1994 sampling (except pJP27 and pJP78). This and the previous sampling more than double the molecular diversity within the crenarchaeotes and demonstrate the power and potential of this approach for microbial ecology and evolution. Given the source, all new isolates must be thermophiles; thus it is especially remarkable that two clones show specific affinity with a marine crenarchaeal sequence SBAR5 PCR amplified from the cold Pacific off the coast of Santa Barbara by Ed DeLong in 1992 (13). The intrusion of such meso- or psychrophilic species takes the "cren" out of the crenarchaeotes [so named in 1990 (7) when all were thought to be like the (presumed) thermophilic ancestor of all Life; cren = spring, fount].

Most of the many analyses Barns and colleagues perform with the addition of these new sequences do show pJP27 and pJP78 as an outgroup to the rest of the Archaea. If this holds, it means (*i*) a demotion in taxonomic rank for Euryarchaeota and Crenarchaeota, the deepest split now separating pJP27/pJP28 (which Barns and coworkers provisionally name Korarchaeota) from all other archaea; (*ii*) that a specific crenarchaeal ("eocyte")/eukaryal affinity is less likely, because the root of the archaeal/eukaryal clade (the branch leading to Bacteria) would have to be moved across two nodes, not one; and (*iii*) that it is anyone's guess what, other than thermophily, pJP27/pJP78 share with other archaebacteria—they could even sport fatty acid glycerol ester lipids or peptidoglycan without violating rules of parsimony.

In fact, they could even *be* eukaryotes: some of the analyses presented have "korarchaeotes" as sister to known members of the Eucarya. The authors do not favor this interpretation, but the mere possibility carries with it the impetus to decide now what we will call new and deeper branches on the line leading to Eucarya, when and if these are ever found, and what precisely we mean by "eukaryote," anyway.

Deconstructing the Eukaryotes

We have seen in our textbooks (and incorporated into our personal biological world views) lists of those fundamental features that distinguish prokaryotes from eukaryotes. Our faith that the eukaryote/prokaryote dichotomy is a natural division in the biological world, a sort of cellular essentialism (14), remains strong in spite of the fact that, individually, many of these distinguishing features (such as mitochondria, Golgi dictyosomes, or 80S ribosomes) turn out to be missing from the most deeply branching eukaryotic lineages, or present in their closest prokaryotic relatives, the Archaea (transcription factors, certain ribosomal proteins). It is as if we believed that the words "eukaryote" and "prokaryote" named natural kinds whose properties we need only to *discover*. But in fact they are categories we ourselves *invented* 30–40 years ago (when our understanding of cell and molecular biology was pretty rudimentary) to define organizational grades or identify evolutionary clades. We have not only the right but the obligation to change them now and in future, as our knowledge grows.

When pressed, most of us would say it is really the nucleus that should make the difference, in fact that "eukaryote" means "true nucleus." But many would then go on to argue that you cannot have a nucleus without an endomembrane system and cytoskeleton, first (15). And no one can really tell us how important the many components of the modern nuclear matrix or envelope were during the evolution of nucleoid to nucleus. Surely complex eukaryotic structural molecules and systems did not appear all at once, and surely we will find homologous genes of similar function when (very soon now) the first archaeal genome sequencing projects start loading data onto the World Wide Web. Already, it is clear that archaeal genomes will contain homologs of genes whose products play important structural roles in the nucleus and cytoplasm of eukaryotic cells—histone, fibrillarin, and tubulins, for instance (16–18).

So, should any new organism that appears on the eukaryote side of the splitting between existing eukaryotes and the Archaea (if they are not paraphyletic) be called a eukaryote? Or should we make a list of all the features that distinguish known eukaryotes from known Archaea and decide among ourselves which or how many are necessary for inclusion in the eukaryote club. Should we be cladists, pheneticists, or evolutionary taxonomists?

What Is an Evolutionary Lineage?

Barns and coworkers' paper touches on a second issue, especially troublesome in the context of other recent analyses focused on protein-coding genes. The rRNA sequences alone, without buttressing arguments based on phenotype, not only fail to support statistically persuasive conclusions about the position of the Korarchaeota but also show some ambivalence about the monophyly of the Archaea (although there is *no doubt* whatever, with the assumed root, that the Archaea and Eucarya form a clade). Such statistical ambivalence is not unusual.

Increasingly, cellular evolutionists seek resolution at the very limits of the power of their algorithms, and it is not clear where greater certainty will come from. Genes for rRNAs are but a tiny fraction of most genomes, of course, so one might hope that parallel work with protein-coding sequences would allow some kind of statistically sensible "meta-analysis." But here we can find disagreement between data sets.

For instance, Baldauf *et al.* (19) have recently completed a study of the EF genes (perhaps the largest relevant assemblage of protein-coding genes). Their analysis, although again overwhelmingly endorsing the archaeal/eukaryal clade, gives support (albeit not strong support) to the sisterhood of crenarchaeotes and eukaryotes—probably the core claim of Lake's eocyte notion. [Given this result, EF sequences from pJP27 and pJP78 would be enormously useful additions to the set, but isolation of organisms corresponding to individual environmental DNA clones remains chancy (20), and walking on mixed clone libraries (21) would be impossibly laborious.]

Many have come to accept that, at least for Eucarya, no single gene can tell the whole phylogenetic story, indeed that there is no single story. Increasingly, we see claims that much of the eukaryotic nuclear genome is of eubacterial origin, either because of (*i*) early and unexpectedly extensive transfer from the proto-mitochondrial endosymbiont (22), (*ii*) frequent independent events of "horizontal gene transfer" (23), or (*iii*) some cataclysmic fusion of archaeal and eubacterial cells and genomes at the founding of the eukaryotes (24). The integrity of archael genomic lineages has not been so seriously impugned, but intermixing with Gram-positive eubacterial genes has been inferred in the case of hsp70 and glutamine synthetase (2, 24, 25).

Thus when the genomic data come flooding in, we may have to decide which genes mark the "true history" of a genomic lineage. I suspect we will settle either for (*i*) "majority-rule," arguing that *most* of the genes in a genome share a common history, which should be taken as the history of the organisms which now bear them, or (*ii*) a core gene (or genes) which functions closest to what we see as the heart of cell biology. Of course, rRNA would be the hands-down favorite of molecular biologists, but a good case could be made for RNA polymerases, or translation EFs.

The "majority-rule" and "core function" approaches both seem arbitrary, and tinged by the same sort of essentialism that colors our thinking about "eukaryotes" and "prokaryotes." We want to believe that organismal and species lineages do have discreet and definable histories that we can discover, and not that we are choosing, arbitrarily, genes whose phylogeny we will equate with that history.

Of course we have already deconstructed the concept of organismal history at the intraspecific level. No sexually reproducing organism is the descendant of a single parent, and vital roles for interstrain and interspecies gene transfer in eubacterial evolution are obvious in several cases. Detectable instances of lateral gene transfer may not be just occasional accidents, important only because they confuse our pictures of true species history, but in fact consequences of the operation of a vital evolutionary mechanism. New evolutionary opportunities might often be met most easily by the replacement of entire genes with homologs having substantially different performance characteristics, rather than through mutation-by-mutation alteration of alleles already in a population. The countervailing forces here would be (*i*) the increasing difficulty of gene transfer and recombination as evolutionary distance between donor and recipient increases, (*ii*) the increasingly radical (and thus sometimes beneficial) differences the imported gene can make to organismal biology, and (*iii*) the extent to which such radical change not only fosters spread of the invading allele within the recipient population by "normal" sexual processes, but protects invaded clades from extinction.

The integrity of organismal genomic lineage surely has been violated by gene transfers, endosymbioses, and "genome fusions," large and small in their consequence. So we can never have a truly cladistic molecular systematics of species, without assumptions about majority-rule or core function which, however generally acceptable, compromise the intellectual purity of the exercise. What in fact we are doing is constructing gene trees, which for various periods of history and at various scales of resolution, have congruent topologies. As we try to go further back in time, or to understand more "primitive" (in terms of cell structure and mechanisms of reproductive isolation) life forms, the less congruence we can expect to find. No single philosophy of systematics will give us the "right" answer about species history because there is no such right answer. But there will be reasonable compromises and generalizations that allow us to talk usefully about the history of life on this planet.

1. Mayr, E. (1981) *Science* **214,** 510–516.
2. Zuckerkandl, E. & Pauling, L. (1965) *J. Theor. Biol.* **8,** 357–366.
3. Schwartz, R. M. & Dayhoff, M. O. (1978) *Science* **199,** 395–403.
4. Woese, C. R. (1987) *Microbiol. Rev.* **5,** 221–271.
5. Kates, M., Kushner, D. J. & Matheson, A. T. (1993) *The Biochemistry of Archaea (Archaeobacteria)* (Elsevier, Amsterdam).
6. Lake, J. A. (1988) *Nature (London)* **331,** 184–186.
7. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.
8. Rivera, M. C. & Lake, J. A. (1992) *Science* **243,** 75–77.
9. Keeling, P. J. & Doolittle, W. F. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 5761–5764.
10. Pace, N. R., Stahl, D. A., Lane, D. J. & Olsen, G. J. (1986) *Adv. Microb. Ecol.* **9,** 1–55.
11. Barns, S. M., Fundyga, R. E., Jeffries, M. W. & Pace, N. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1609–1613.
12. Barns, S. M, Delwiche, C. F. Palmer, J. R. & Pace, N. R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 9188–9193.
13. DeLong, E. F. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 5685–5689.
14. Sober, E. (1980) *Philos. Sci.* **47,** 350–383.
15. Cavalier-Smith, T. (1991) in *Foundation of Medical Cell Biology,* ed. Bittar, G. E. (J.A.I. Press, Greenwich, CT), Vol. 1, pp. 217–272.
16. Starich, M. R., Sandman, K., Reeve, J. N. & Summers, M. F. (1995) *J. Mol. Biol.* **255,** 187–203.
17. Potter, S., Durovic, P. & Dennis, P. P. (1995) *Science* **268,** 1056–1057.
18. Margolin, W., Wang, R. & Kumar, M. (1996) *J. Bacteriol.* **178,** 1320–1327.
19. Baldauf, S. L., Palmer, J. D. & Doolittle, W. F. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7749–7754.
20. Huber, R., Burggraf, S., Mayer, T., Barns, S., Rossnagel, P. & Stetter, K. O. (1995) *Nature (London)* **376,** 57–58.
21. Stein, J. L., Marsh, T. L, Wu, K.-Y., Shizua, H. & DeLong, E. F. (1996) *J. Bacteriol.* **178,** 591–599.
22. Clark, C. G. & Roger, A. J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6518–6521.
23. Smith, M. W. & Doolittle, R. F. (1992) *Trends Biochem. Sci.* **17,** 489–493.
24. Golding, G. B. & Gupta, R. S. (1995) *Mol. Biol. Evol.* **12,** 1–6.
25. Brown, J. R., Masuchi, Y., Robb, F. T. & Doolittle, W. F. (1994) *J. Mol. Evol.* **38,** 560–576.