The nature of the universal ancestor and the evolution of the proteome

W Ford Doolittle

The past year has seen several attempts to reconstruct the proteome of the universal ancestor of all life on the basis of comparisons of contempory genomes. However, increasing evidence for lateral gene transfer could mean that such attempts are based on an incorrect understanding of evolution.

Addresses
Canadian Institute for Advanced Research, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada; e-mail: ford@is.dal.ca

Current Opinion in Structural Biology 2000, **10**:355–358

0959-440X/00/$ – see front matter

Abbreviation
LGT lateral gene transfer

Introduction

Near the end of *The Origin of Species*, Darwin recapitulates his argument that patterns of similarity and difference among organisms reflect their descent with modification from an ever smaller number of ever more ancient ancestors. Extending this reasoning to its logical end point, he infers "that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed" [1].

Whether 'one primordial form' denoted a single cell or a single species, it is clear that Darwin envisioned the universal ancestor as an entity with a uniquely definable phenotype, however primitive. Most contemporary theorists reason in the same way, although disagreeing about the nature and complexity of the ancestral phenotype. The enzymes of metabolism and the proteins involved in the replication and expression of genes (and of course the code) are just too similar among all known species to be of independent nonliving origin. The conclusion that all contemporary organisms must have derived from a single 'form', in whose genome the ancestral versions of all these proteins were encoded, seems inescapable.

Can we reconstruct the proteome of that universal ancestor? In recent years, many authors [2,3•,4,5,6•,7–10] have used comparisons of the gene contents of modern genomes in attempts to do just that. In fact, the November 1999 issue of the *Journal of Molecular Evolution* is entirely dedicated to the understanding of the biology of the universal ancestor (which some call 'the last universal common ancestor' or 'LUCA', and some call 'the cenancestor'). Other researchers, however, have cast the rooting of the universal tree on which such analyses ultimately depend into doubt [11,12,13••] or have claimed that lateral transfer of genes between species, phyla or domains is so frequent that all reconstruction attempts are doomed to failure [14,15••]. Most radical among these, Woese [16••] has argued that there never was a universal ancestral cell or species, but rather that a complex population of heterogeneous genetic entities — 'progenotes' — gave rise to modern cellular lineages. This view, now increasingly supportable, is a profound challenge to our understanding of genome origins and the evolution of the proteome.

Did the universal ancestor have a large and 'modern' genome?

The widely endorsed universal tree of life recognizes three primary domains (bacteria, archaea and eukarya), first defined by sequences of SSU (small subunit) rRNAs. The most often accepted rooting of this tree, based on paralogous protein-coding gene families [17•], has its deepest division (earliest branching) separating bacteria, on the one side, from a lineage that later diverged into archaea and eukaryotes, on the other. Given this tree, we should be able to use reasoning based on parsimony to infer the composition of the genome of the universal ancestor. Parsimony tells us that — *barring lateral gene transfer* (LGT) between species — any gene present in organisms on both sides of the deepest branching was probably present in the universal ancestor [18]. Otherwise, its appearance on either side of this division would require two independent inventions and (if the genes show significant similarity) unlikely sequence convergence. (Note that this argument retains its force even when only one bacterium and one archaean have the gene — as long as we disallow LGT.) Genes present only on one side, that is only in bacteria or only in archaea plus eukarya, could be either recent ('invented' on that side) or ancestral (lost on the other). Genes restricted to a group of related organisms within a domain are most likely to be recently invented.

Nature is not obliged to behave parsimoniously: sometimes independently invented genes *will* converge in parts of their sequence and (more often) independent losses of the same ancestral genes in all but a single lineage will give patterns mimicking recent origin. Nevertheless, parsimony is the only logical guide we have. In the universal ancestor issue of the *Journal of Molecular Evolution*, several authors apply such reasoning to the question of the genome/proteome of that ancient cell. Castresana and Moreiera [8], for instance, infer from comparing bacterial and archaeal sequences in the databases that the universal ancestor enjoyed the use of "at least four electron transport chains [oxygen, nitrate, sulfate and sulfur respiration], and therefore...may have been prepared to face a wide range of environmental conditions". Similarly, Labedan *et al.* [9] conclude that "the last common ancestor to all extant life possessed differentiated [multiple] copies of genes coding

for both [ornithine and aspartate] carbamoyltransferases, indicating it as a rather sophisticated organism". Kyrpides *et al*. [7] derive similar results from a more exhaustive analysis, addressing all the genes in the *Methanococcus jannaschii* genome. They conclude that the universal ancestor "contained metabolic enzymes and genetic systems similar to those of extant unicellular organisms".

Other recently published complete genome analyses, although not specifically addressing the genome/proteome of the universal ancestor, only add to the tally of genes that are present in both bacteria and archaea, and thus are attributable to the universal ancestor — if we don't allow LGT. For instance, Koonin and collaborators [19••] have just presented a study of four sequenced euryarchaeal genomes that shows that only 31–35% of each genome comprises core genes shared by all four species: the clear majority are made up of genes found in only some or none of the other species and the clear majority of these have obvious homologs within bacteria. In a similar analysis of four euryarchaeal and one crenarchaeal genome, Faguy and I [20] found that between 6 and 13% of each genome comprises 'bacterial genes' not found in the other four. Thus, there is a very large pool of genes shared among contemporary bacteria and archaea that are never all (or even mostly) found in any one bacterial or archaeal genome. If we follow the simple rules of parsimony without LGT, we wind up with a *totipotent ancestor*, with a proteome considerably more complex than that of any modern prokaryote. This cell must have been capable of almost the full range of the autotrophic and heterotrophic, anaerobic and aerobic biochemistries separately found among all the diverse prokaryotes in today's microbiota!

Was the universal ancestor a eukaryote?

One way of resolving this might be to suggest that the universal ancestor was not a typical prokaryote. Forterre and Philippe [13••] have recently presented a scheme in which a gene-rich ancestral genome has a somewhat more comfortable place. They note that the accepted paralog-based rootings of the universal tree could all be erroneous (artifacts of 'long-branch attraction' or unrecognized deeper paralogy): the true root could as easily separate eukaryotes on the one hand from archaea and bacteria on the other. Thus, the ancestor might well have been a complex cell like a eukaryote, with archaea and then bacteria having evolved from it by genome reduction and streamlining (possibly in adapting to high temperature). Penny and colleagues [21,22,23•] have added a twist of parsimony to this line of thought. The RNA world (if it existed) is the outgroup to all cellular life. If the RNAs used in intron splicing and stable RNA processing in modern eukaryotes are relics of the RNA world, then they must have been present in the universal ancestor. Thus, that ancestor was itself a eukaryote (at least in these respects) and archaea and bacteria represent successive stages in the replacement of RNA catalysis by protein catalysis. This idea is not a new one: Hartman, Darnell and I (and probably others) [24–26] put forth versions of it more than 20 years ago, in conjunction with the hypothesis that introns are relics of some precellular gene assembly process.

Is a single universal ancestral cell or species really necessary?

Even earlier than that, Woese and Fox [27,28] described quite a different way to avoid the totipotent gene-rich universal ancestor — basically by embracing LGT as the key feature for understanding early cell evolution. They considered that the three contemporary domains of life arose not from a single cell, but from a population of very different cellular entities ('progenotes') that were primitive in two respects. First, their machineries of replication, transcription and translation were, as yet, inefficient and inaccurate — so their genomes had to be small to avoid error catastrophe. The population as a whole might have contained ancestral forms of all the genes in the large pool of genes now shared among contemporary bacteria and archaea, but no single member of it did. Second, they were promiscuous participants in LGT. In a recent and more thorough articulation of this concept, stimulated by the growing appreciation of the evolutionary importance of LGT even among modern prokaryotes, Woese [16••] describes the progenote stage as follows: "Their small genomes require progenotes to be metabolically simple, minimal. However, different progenotes could have differed metabolically. The [collective] genetic complement of the progenote population could have been far greater than that of any individual cell, indeed totipotent... The fact that innovations could easily spread through the population by lateral gene transfer gave the progenote community enormous evolutionary potential ...".

How could such a population give rise to two (and then three) discrete cellular domains without passing through a bottleneck represented by a single cellular universal ancestor? Perhaps in almost the same way that sexually reproducing species speciate, giving rise to daughter species whose gene pools are initially similar to those of the parent species and contain very many more different alleles than are borne by any one genome. (What is crucially not the same is that the genomes of the progenote population — like those of different modern prokaryotes — bore many different genes, not just different alleles.) Thus, there is no more reason to imagine only a single first kind of cell as the progenitor of all contemporary life than there is to imagine only Adam and Eve as progenitors of the human species.

Still, one might object, the various different genes shared among the members of the heterogeneous progenote population Woese envisions did not suddenly appear from nowhere, but themselves had ancestors. Could not the ancestral versions of these different genes have resided together in a single genome — which we could then call the universal ancestor — in some even more distant past? If by ancestral versions of genes we mean the founders of

today's protein-coding superfamilies or indeed the subgenic modules encoding the 1000 or so protein structural motifs thought to be around now [29], this genome might not have to have been any larger than that of a modern prokaryote. Or, even if at this stage there was still a heterogeneous population of genomic entities, might not the earlier triumph of the modern universal genetic code have represented the bottleneck event to which all surviving coding regions can be traced? Or, failing even that, could not the RNA world antecedents of protein-motif-encoding DNAs be traceable to a single first self-replicating RNA, the one true universal ancestor?

These all seem like sensible possibilities, but there is no reason to suppose that any of them relate to any ancestral state that we might reconstruct by looking at the distribution of genes among contemporary species. Each family of related genes should, in principle, be traceable to a single last common ancestral version; however, the ancestral versions of different families will have existed in different genomes at different times during life's history. To use the analogy to human evolution again, all human mitochondrial DNAs (barring recombination among them) can trace their ancestry to a single mitochondrial DNA in a unique ancient woman, popularly called 'mitochondrial Eve' [30]. However, other genes in contemporary human genomes are derived from common ancestral genes in the genomes of different members of the human or prehuman population, living before or after that woman's time. Mitochondrial Eve surely never met Y-chromosomal Adam!

There is also not much reason to suppose that the last common ancestors of any contemporary gene families inhabited cells that were substantially different from modern cells. Although transcription and translation machineries clearly have domain-specific features, most of the components involved are homologous. Presumably, these functions were reasonably modern before the time of the ancestral versions of the genes encoding these components. The strongest case for divergence from a primitive state has been made for the DNA replication machinery. As Liepe *et al.* [31••] recently summarized the situation, bacterial replicative polymerases and primases are clearly nonhomologous among bacteria and archaea/eukaryotes, whereas the principal replicative helicases and proof-reading exonucleases contain both homologous and nonhomologous domains. The sliding clamp proteins and DNA ligases, though homologous, are highly diverged. Only a few replication components (clamp-loader ATPases and 5′→3′ exonucleases) and enzymes of DNA precursor metabolism and manipulation (topoisomerases and gyrases) are well conserved across the tree's deepest branching. In 1996, Mushegian and Koonin [32] interpreted these observations to mean that the universal ancestor had an RNA genome — DNA replication systems evolving quasi-independently after the universal tree's first branching. Now, Koonin (with Liepe and Aravind) [31••] articulates a subtler scenario, in which the universal ancestor had an RNA genome that replicated through cycles of reverse transcription, RNase H digestion, ssDNA-templated dsDNA synthesis and transcription.

Even in this case, however, the occurrence of LGT allows a radically different view. Forterre [33], impressed by the fact that a few replication components *are* homologous and well conserved across domains, argues that the other components differ because they have been replaced (probably on the bacterial side) by functionally analogous genes from plasmids and phage. Certainly, the presence of such extrachromosomal 'selfish DNAs' provides an environment for the rapid evolutionary diversification of replication and segregation functions, separate from, but in contact with, bacterial genomes. Typically, extrachromosomal DNAs do not encode transcription and translation components: hence, the relative evolutionary conservatism of these functions across domains.

Life without a cellular ancestor: implications for the concept of homology

Evolutionary theory now figures prominently in the thinking and writings of molecular geneticists and structural biologists. In particular, the understanding that 'homology' denotes descent from a common ancestor rather than sequence similarity — although sequence similarity can be taken as evidence for homology — is now quite general [34]. Homology is a matter of quality, not quantity and the oxymoronic term 'percent homology' is seldom seen these days.

However, homology is still a funny word: in the context of proteins and genes, it makes sense only if we don't think about it too deeply. If our model of evolution invokes a single cell as the universal ancestor, then we might conveniently trace all modern genes back to one or another particular 'family founder' gene in the universal ancestral genome. Genes are thus homologous if and only if they are members of such a family. I think that many discussions of protein families and superfamilies embrace such a concept, at least implicitly [35,36]. Genes in the universal ancestor that were already homologs of each other (the paralogs such as elongation factors EF-1α and EF-G used to root the universal tree, for instance [37]) of course complicate this view. Still, the universal ancestor seems to represent a sort of horizon beyond which we can justify not looking.

If there was no ancestor, however, how can we avoid thinking about the possibility that all genes are ultimately derived from a single short RNA, the first replicating ribozyme. If this is true, *all genes are homologous*. We might still be able to distinguish between orthologs and paralogs, as a matter of logical principle but, in practice, this will often be impossible. 'Homology' itself becomes a useless word unless we redefine it to mean something like 'statistically more similarity than we would expect on the basis of chance'. Such an operational definition is slippery — genes can fade in and out of a state of homology depending on the kinds of analysis and the

background database within which we compare them. It's a short step from here back to 'percent homology'. It is ironic that the words we seem to need in order to think productively about biology, words such as 'homology', 'individual', 'organism' and 'species', have no precise meaning [38].

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Darwin C: *The Origin of Species by Means of Natural Selection*. London: J Murray; 1859.

2. Ouzounis C, Kyrpides N: **The emergence of major cellular processes in evolution.** *FEBS Lett* 1996, **390**:119-123.

3. • Aravind L, Walker RW, Koonin EV: **Conserved domains in DNA repair proteins and evolution of repair systems.** *Nucleic Acids Res* 1999, **27**:1224-1242.
The authors used parsimony to conclude that the universal ancestor encoded a RecA-like recombinase and a variety of other repair functions.

4. Kyrpides N, Woese CR: **Universally conserved translation initiation factors.** *Proc Natl Acad Sci USA* 1998, **95**:224-228.

5. Tomii K, Kanehisa M: **A comparative analysis of ABC transporters in complete microbial genomes.** *Genome Res* 1998, **8**:1048-1059.

6. • Lazcano A, Forterre P: **The molecular search for the last common ancestor.** *J Mol Evol* 1999, **49**:411-412.
This editorial sets the stage for a special issue on the universal ancestor that comprises the (generally updated) proceedings of a conference held in France in 1996.

7. Kyrpides N, Overbeek R, Ouzounis C: **Universal protein families and the functional content of the last universal common ancestor.** *J Mol Evol* 1999, **49**:413-423.

8. Castresana J, Moreiera D: **Respiratory chains in the last common ancestor of living organisms.** *J Mol Evol* 1999, **49**:453-460.

9. Labedan B, Boyen A, Baetens M, Charlier D, Chen P, Cunin R, Durbeco V, Glansdorff N, Hreve G, Legrain C *et al.*: **The evolutionary history of carbamoyltransferases: a complex set of paralogous genes was already present in the last universal common ancestor.** *J Mol Evol* 1999, **49**:461-473.

10. Diruggiero J, Brown JR, Bogert AP, Robb FT: **DNA repair systems in archaea: momentos from the last universal common ancestor?** *J Mol Evol* 1999, **49**:474-484.

11. Philippe H, Forterre P: **The rooting of the universal tree is not reliable.** *J Mol Evol* 1999, **49**:509-523.

12. Brinkmann H, Philippe H: **Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies.** *Mol Biol Evol* 1999, **16**:817-825.

13. •• Forterre P, Philippe H: **Where is the root of the universal tree of life?** *BioEssays* 1999, **21**:871-879.
The authors summarize the various reasons for placing little faith in deep phylogenetic trees. Suitably corrected data indicate that the root of the universal tree should be between eukaryotes on one side and bacteria plus archaea (prokaryotes) on the other.

14. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284**:2124-2128.

15. •• Martin W: **Mosaic bacterial chromosomes: a challenge en route to a tree of genomes.** *BioEssays* 1999, **21**:99-104.
An excellent summary of the case for and implications of lateral gene transfer in prokaryotes.

16. •• Woese CR: **The universal ancestor.** *Proc Natl Acad Sci USA* 1998, **95**:6854-6859.
An eloquent and well-reasoned explanation of Woese's 'progenote hypothesis', which appeared in inchoate form in [27]. This paper makes a compelling case for the notion that there was no universal ancestral organism and contains many reasoned speculations about the evolution of cellular complexity.

17. • Gribaldo S, Cammarano P: **The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery.** *J Mol Evol* 1998, **47**:508-516.
The most recent of several papers using paralogous gene families to root the universal tree. Most published studies favor a 'bacterial rooting', with the deepest division separating bacteria from a line that later diverged into archaea and eukaryotes. Philippe and Forterre [11,12,13••] believe this result to be artifactual.

18. Harvey PH, Pagel MD: *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press; 1991.

19. •• Makarova KS, Aravind L, Galperin MY, Grishin NV, Tausov RL, Wolf YI, Koonin EV: **Comparative genomics of the archaea (euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell.** *Genome Res* 1999, **9**:608-628.
A comparison of four completely sequenced euryarchaeal genomes shows a surprisingly small set of universally conserved genes, which are mostly involved in replication, transcription or translation. Many genes appear to have been laterally transferred from bacteria.

20. Faguy DM, Doolittle WF: **Lessons from the *Aeropyrum pernix* genome.** *Curr Biol* 1999, **9**:R883-R886.

21. Poole AM, Jeffares DC, Penny D: **The path from the RNA world.** *J Mol Evol* 1998, **46**:1-17.

22. Poole AM, Jeffares DC, Penny D: **Early evolution: prokaryotes, the new kids on the block.** *BioEssays* 1999, **21**:880-889.

23. • Penny D, Poole A: **The nature of the universal common ancestor.** *Curr Opin Genet Dev* 1999, **9**:672-677.
The most recent of several publications in which Penny and Poole argue that the universal ancestor was a eukaryote.

24. Hartman H: **The origin of the eukaryotic cell.** *Spec Sci Technol* 1984, **7**:77-81.

25. Darnell JE Jr: **Implications of RNA-RNA splicing in evolution of eukaryotic cells.** *Science* 1978, **202**:1257-1260.

26. Doolittle WF: **Genes-in-pieces: were they ever together?** *Nature* 1978, **272**:581-582.

27. Woese CR, Fox GE: **The concept of cellular evolution.** *J Mol Evol* 1977, **10**:1-6.

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

29. Wolf YI, Brenner SE, Bash PA, Koonin EV: **Distribution of protein folds in the three superkingdoms of life.** *Genome Res* 1999, **9**:17-26.

30. Gibbons A: **Mitochondrial Eve: wounded, but not dead yet.** *Science* 1992, **257**:873-875.

31. •• Liepe DD, Aravind L, Koonin EV: **Did DNA replication evolve twice independently?** *Nucleic Acids Res* 1999, **27**:3389-3401.
A thorough and reasoned comparative analysis of the components of DNA replication in life's three domains. The authors conclude that the universal ancestor replicated DNA differently compared with many modern organisms.

32. Mushegian AR, Koonin EV: **A minimal gene set for cellular life derived by comparison of complete bacterial genomes.** *Proc Natl Acad Sci USA* 1996, **93**:10268-10273.

33. Forterre P: **Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins.** *Mol Microbiol* 1999, **33**:457-465

34. Reeck GR, de Haen C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH *et al.*: **'Homology' in proteins and nucleic acids: a terminology muddle and a way out of it.** *Cell* 1987, **50**:667.

35. Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM: **Ancient conserved regions in new gene sequences and the protein databases.** *Science* 1993, **259**:1711-1716.

36. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**:631-637.

37. Baldauf SL, Palmer JD, Doolittle WF: **The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny.** *Proc Natl Acad Sci USA* 1996, **93**:7749-7754.

38. Fox Keller E, Lloyd EA (Eds): *Keywords in Evolutionary Biology*. Cambridge, Massachusetts: Harvard University Press; 1992.