Cell, Vol. 89, 995–998, June 27, 1997, Copyright ©1997 by Cell Press

Archaea and the Origin(s) of DNA Replication Proteins

Minireview

David R. Edgell and W. Ford Doolittle
Canadian Institute for Advanced Research
Department of Biochemistry
Dalhousie University
Halifax, Nova Scotia B3H 4H7
Canada

The deepest phylogenetic division in the universal tree (vertical line in Figure 1) is that separating bacteria from the clade comprising archaea and eukaryotes. The prokaryote–eukaryote split (horizontal line in Figure 1), originally delineated on the basis of differences between (eu)bacterial and eukaryotic cellular ultrastructure, is a phenetic dichotomy (see Doolittle, 1996, and references therein). How useful this dichotomy still is remains an active question: its answer will come from a full appreciation of molecular and cellular differences between archaea and eukaryotes. Part of the answer seems to be known. In many ways, the transcription, translation, and splicing machineries of archaea look very eukaryotic (see minireviews [this issue of *Cell*] by Belfort and Weiner, Dennis, and Reeve et al.). Much of the reworking of the gene-expression apparatus we previously considered part of the "prokaryote-to-eukaryote transition" actually occurred earlier than that transition, before the archaeal–eukaryotic divergence.

What about replication? Archaeal genomes seem distinctly bacterial in character. Archaea have single circular chromosomes on which genes are tightly packed, sometimes overlapped and often linked in operons. Some archaeal operons are identical in gene order to bacterial operons and must have existed in this form since the time of the cenancestor. The closest archaeal relative of eukaryotic tubulin (whose appearance was essential for the evolution of mitosis) is the FtsZ protein, but it much more strongly resembles bacterial FtsZ proteins than tubulin in structure and apparent function (Margolin et al., 1996). If substantial changes in the replication machinery accompanied the change from prokaryotic to eukaryotic chromosome structure, replication pattern (one origin to many), and segregation mechanism (mitosis), then we would expect these changes to have occurred in the eukaryotic lineage after it diverged from archaea. We would expect archaea to look like bacteria in terms of replication proteins.

Surprisingly, Archaeal Replication Proteins Look More Like Eukaryotic Replication Proteins

Even before the appearance of significant archaeal genome-sequence data, one could begin to see that this expectation might not be met. Early studies of DNA replication in vivo demonstrated that halophilic archaea were sensitive to aphidicolin, a specific inhibitor of eukaryotic but not bacterial replicative DNA polymerases (reviewed in Forterre and Elie, 1993). Every archaeal DNA polymerase sequenced to date is a eukaryote-like family B DNA polymerase, and the complete genome sequence of Methanococcus jannaschii reveals only a single (family B) DNA polymerase (Bult et al., 1996), making it likely that it is the replicative enzyme.

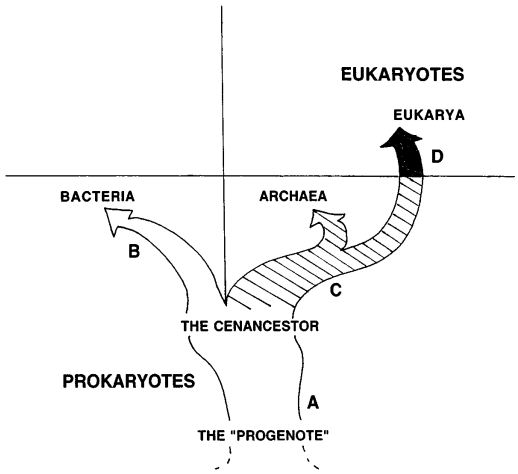

Figure 1. Structure of the Universal Tree

The vertical line indicates the deepest known phylogenetic split, which separates Bacteria from the lineage that gave rise to Archaea and Eucarya (Gogarten et al., 1989; Iwabe et al., 1989). The horizontal line represents the division between prokaryote and eukaryotes, which is a phenetic one, based primarily on cellular ultrastructure (for review, see Doolittle, 1996). The cenancestor is the last common ancestor of all living organisms. The progenote is an hypothetical (but logically necessary) ancestor in which basic information-handling processes (replication, transcription, and translation) were still undergoing rapid and fundamental evolutionary change. Application of principle of parsimony to the distribution of characteristic components of major cellular processes among domains suggests: (i) that transcription and translation had evolved by point (A), and DNA genomes with operons (some with modern gene order) were present by that time; (ii) that the eukaryote-like features of archaeal transcription, translation, and replication discussed in this review and others in this issue either evolved along the branch designated by (C) or else appeared at (A) and then were lost or radically altered in the Bacteria (B); and (iii) the distinguishing features of cell ultrastructure that underpin the "prokaryote–eukaryote dichotomy" first appeared at point (D).

Moreover, the M. jannaschii genome sequence (and scattered individual archaeal gene sequences) shows many ORFs that are clearly most closely (or only) related in sequence to eukaryotic replication proteins (Bult et al., 1996). There are three general categories of results from sequence comparisons of replication proteins of bacteria, archaea, and eukaryotes. First, there are archaeal ORFs with clear evidence of homology to eukaryotic replication proteins, while evidence that either the archaeal or eukaryotic protein shares a common ancestor with bacterial proteins performing the same function is weak or absent. Several such "eukaryote-specific" proteins are listed in Table 1. Second are situations in which bacterial and eukaryotic replication proteins are likely homologs, but the archaeal and eukaryotic versions are by far the more similar. For instance, Methanococcus possesses two homologs of the clamp-loading complex that are more similar at the amino acid level to eukaryotic clamp-loading proteins (replication factor C; [RFC]) than to the bacterial homologs (the DnaX and HolB proteins). The third sort of result involves replication functions performed by a number of homologous

Table 1. Similarities of Bacterial, Archaeal, and Eukaryotic Replication Proteins

Function at replication fork	Eubacterial protein (E. coli)	Eukaryotic protein (yeast/human)	Archaeal protein[1]	Evidence for homology[2]
origin recognition	DnaA	Origin recognition complex (ORC) proteins 1-6	ORC1-like (plasmid encoded) M. thermoformicicum	little or no 1° sequence similarity between eub and euk/arch; all ATP-dependent[3]
single-strand DNA-binding protein; loading of helicase, stimulates DNA polymerase	SSB	Replication protein A (RPA; 3 subunits)	?????[4]	little or no 1° sequence similarity between eub/euk; 3° structure similarity in SSB-binding domain (see text)
synthesis of primer	DnaG	DNA polymerase α (Family B DNA polymerase)[5]	Family B DNA polymerase (many archaeal sequences)	little or no 1° sequence similarity between DnaG and DNA polymerase α[6]
helicase	DnaB (5'-3' helicase) PriA (3'-5' helicase)	Dna2 (3'-5' helicase)	archaeal Dna2	little or no 1° sequence similarity among helicases, expept in ATP-binding pocket; different activities and substrate specificities[7]
clamp loading complex; DNA-dependent ATPase, stimulates loading of DNA polymerase	γ complex (γδδ′χΨ) **dnaX=γ subunit holB=δ′ subunit**	**Replication factor-C (RFC) 5 homologous subunits, RFC 1-5**	**archaeal RFC-1 archaeal RFC-3**	significant a.a. similarity between eub/euk/arch; similar biochemistry (see text)
processivity factor, 'sliding clamp'	Polβ (dnaN)	Proliferating cell nuclear antigen (PCNA)	archaeal PCNA	little or no 1° sequence similarity but crystal structures of eub & euk proteins are identical and superimposable; see text
synthesis of leading and lagging strands	DNA polymerase III core (αθε) (Family C DNA polymerase)	DNA polymerase α/ε/δ (Family B DNA polymerase)	Family B DNA polymerase (many archaeal sequences)	little or no 1° sequence similarity between different families of polymerases; all have similar biochemical activities (3'-5' exo, polymerization)[8]
ligation of fragments on lagging strand	DNA ligase (NAD-dependent)	DNA ligase (ATP-dependent)	DNA ligase (ATP-dependent) M. jannaschii, D. ambivalens, M. thermoformicicum	little or no 1° sequence similarity and different nucleotide co-factors
removal of primers	DNA polymerase I (Family A DNA polymerase); **Ribonuclease H**	FEN1/Rad2 (pombe); **Ribonuclease H**	archaeal FEN1/Rad2; **Ribonuclease H**	ribonuclease H's homologous; the 5'-3' exo domain of E. coli DNA polI and FEN1/Rad2 are claimed to be homologous but alignments are weak.[9]

All eukaryotic and archaeal replication proteins share significant amino acid similarity. None of the bacterial replication proteins share signficant similarity with either eukaryotic or archaeal proteins performing analogous functions except those that are boxed. Table is based on Stillman, 1994.

[1] Archaeal proteins are from M. jannaschii unless indicated.

[2] Evolutionary biologists use the term homology to refer to the historical relationship of (for instance) two or more proteins: two proteins are homologous if they evolved by descent from a common ancestral sequence (Reeck et al., 1987). Amino acid sequence similarity is often the only criterion for judging homology; a common function alone is not sufficient evidence for homology because two proteins can convergently (and independently) arrive at the same mechanistic, structural, or biochemical solution to a particular biological problem. Similarity refers to conserved amino acid substitutions (i.e., Ile→Val, Trp→Phe), while identity refers to the same amino acid in homologous positions. Proteins that share a significant amino acid identity (usually 20%–25%, with allowance for gaps) are considered to be homologs (Doolittle, 1986). Two or more proteins with less than this level of sequence identity (which is considered no better than a random alignment of two amino acid sequences) might be homologs and may have evolved from a common ancestral sequence but have diverged too much in sequence to allow reconstruction of their relationship.

[3] In addition to limited sequence similarity, DnaA and ORC proteins exhibit a number of functional differences. ORC proteins are constitutively bound to ARS sequences in yeast throughout the cell cycle (for review, see Diffley, 1996), whereas in E. coli DnaA is prevented from binding *oriC* because *oriC* is bound by the SeqA protein, which acts negatively to regulate DNA-replication initiation.

[4] Question mark indicates that no predicted open reading frame from the M. jannaschii genome with significant similarity to known single-strand DNA-binding proteins was found.

[5] DNA-dependent DNA polymerases are classified into families based on amino acid similarity to one of the three E. coli DNA polymerases (Braithwaite and Ito, 1993). Family A DNA polymerases are similar to E. coli DNA polymerase I (pol A), family B DNA polymerases are similar to E. coli DNA polymerase II (pol B), and family C DNA polymerases are similar to E. coli DNA polymerase III (pol C). DNA polymerases of different families cannot be aligned on the amino acid level with any confidence.

[6] The eubacterial primase, DnaG, and eukaryotic DNA polymerase α are claimed to have homologous functional residues in conserved domains (Figure 4 of Prasartkaew et al., 1996). However, only 4 of 19 (21%) residues of E. coli DnaG and Homo sapiens DNA polymerase α are similar in motif A; none are identical. Of 16 residues of motif C, only 1 is identical (6%), while 3 are similar (18%). The proteins are not alignable outside of these domains.

[7] Identification of a eukaryotic replication fork–associated helicase has been problematic. Dna2, a yeast helicase, associates with the 5′–3′exo–endonuclease FEN1/Rad2 (pombe) of yeast and is most likely involved in Okazaki fragment maturation (Budd and Campbell, 1997). It is not clear if Dna2 is associated with origin unwinding (as is PriA in E. coli) or with unwinding of the replication fork (as is DnaB in E. coli).

[8] Three family B DNA polymerases, α, δ, and ε, have been identified as essential for replication in S. cerevisiae (for review, see Stillman, 1994). The exact biochemical role of DNA polymerase ε in other eukaryotes is not known. It is not required for SV40 replication but may be necessary for replication in mammalian cell lines.

[9] The 5′–3′ exonuclease domain of eubacterial DNA polymerase I and the eukaryotic 5′–3′ exonuclease FEN-1/Rad2 (pombe) have been classified as members of a homologous protein family based on amino acid alignments (reviewed in Lieber, 1997). However, the 5′–3′ exonuclease domain of E. coli DNA polymerase I (301 amino acids) and murine FEN1 (337 amino acids) are only 21% similar.

eukaryotic proteins, which appear to be reduced in number in Methanococcus. Thus, there are three replicative family B DNA polymerases (α, δ, and ϵ) in eukaryotes (Braithwaite and Ito, 1993) but only a single homolog in Methanococcus; five clamp-loading (RFC) proteins in eukaryotes (O'Donnell et al., 1993; Cullman et al., 1995) but only two in Methanococcus; six minichromosome maintenance (MCM) proteins (control of initiation of replication) in eukaryotes (Kearsey et al., 1996) but only three in Methanococcus.

Gaps in the Data

While Methanococcus may have a basic set of eukaryote-like replication proteins, there are a number of critical components that appear to be missing or that have not yet been identified from the complete genome sequence. For instance, no single-strand DNA-binding protein was identified, yet this protein is essential for initiation of replication in both bacteria and eukaryotes (Kornberg and Baker, 1992).

Also critical for initiation of replication are origin-binding proteins, yet neither bacterial nor eukaryotic homologs were reported in the initial publication. Subsequent work by other researchers identified a possible homolog of the bacterial origin-binding protein DnaA (http://www.tigr.org/tdb/mdb/mjdb/updates/update_090596.html). However, it is unlikely that this protein is a true homolog of DnaA, as database searches with this ORF have low significance values, and the predicted protein is a member of the largest gene family ($>$20 genes) in the Methanococcus genome. Two sequences in databases, not from Methanococcus but from the closely related Methanobacterium thermoformicicum, are possible homologs of the ORC1 protein of eukaryotes (Nolling et al., 1992). Curiously, these genes are present on plasmids and may be important for plasmid maintenance and replication. It is not clear what role, if any, these proteins might play in chromosomal replication.

Many Bacterial and Eukaryotic (Archaeal) Replication Proteins Are Not Similar at the Amino Acid Level yet Perform Analogous Functions

Comparisons of the amino acid sequences of the proteins corresponding to analogous activities from bacteria and eukaryotes reveal that many of these proteins are very dissimilar (Table 1); amino acid alignments are often no better than an alignment of two random sequences, and the proteins are often radically different in length and subunit composition. Crystal structures of some bacterial and eukaryotic replication proteins have been solved, and comparisons of these structures can help to address issues of common ancestry. We discuss two examples: single-strand binding proteins and processivity factors.

One of the first steps in the initiation of replication is the binding of the unwound origin region by single-strand DNA-binding protein (Kornberg and Baker, 1992). In E. coli, this function is performed by SSB (single-strand DNA-binding protein) and in eukaryotes, by RPA (replication protein A). RPA is a heterotrimeric complex (in Homo sapiens, 70, 34, and 11 kDa) with the ssDNA-binding activity residing in the large subunit. All of these proteins have been described as homologs, but published amino acid alignments are not compelling (Philipova et al., 1996). The E. coli SSB protein is 177 amino acids in length, yet of those residues only 19% are similar to the (longer) eukaryotic second subunit, while only 16% are similar to the (shorter) third largest subunit; these percentage similarities are no better than random alignments. Furthermore, multiple gaps (indicating many independent insertion and deletion events in the evolution of these genes) must be introduced to align the largest RPA subunit with the E. coli SSB protein in regions of the proteins thought to be essential for SSB activity. Based solely on these amino acid alignments, it is difficult to be convinced that eukaryotic and bacterial SSB proteins are homologs.

Recently, the crystal structure of the SSB-binding domain of human RPA bound to ssDNA was solved (Bochkarev et al., 1997). Although the structures of two other replication-associated SSB-binding domains have also been determined, that of the gene V protein of bacteriophage f1 and the gp32 protein of bacteriophage T4, the RPA SSB domain was most similar to that of S. cerevisiae aspartyl-tRNA synthetase bound to tRNA. This finding does not convincingly show that these replication-associated SSB proteins evolved from a common ancestral DNA-binding protein; other explanations are equally likely. For instance, the ability to bind single-stranded nucleic acids might have evolved independently many times, or an ancestral SSB domain might have been shuffled between proteins that originally lacked this ability.

Less ambiguous is the case of the processivity factor proliferating cell nuclear antigen (PCNA) in eukaryotes and pol β (*dnaN*) in bacteria. These proteins, often called sliding clamps, are responsible for loading the DNA polymerase onto the active template and ensuring processive replication. Both the bacterial and eukaryotic versions have very similar biochemistry and are functionally analogous. The amino acid sequences of the eukaryotic and bacterial proteins are not significantly similar (amino acid similarity is below that of a random alignment), but the crystal structures of the intact bacterial and eukaryotic proteins are almost identical and can be superimposed (Kelman and O'Donnell, 1995). Given that the structural similarity extends over the entire length of the proteins rather than being confined to a single functional domain, it is likely that these two proteins did indeed evolve from a common ancestral sliding-clamp protein and have since diverged in sequence.

The Cenancestor Had a DNA-Based Genome

It is surprising that the replication proteins of bacteria and archaea–eukaryotes show little or no sequence similarity. Core protein components of other genetic processes (such as DNA-dependent RNA polymerases, elongation factors, initiation factors, and some ribosomal proteins) share significant sequence similarity, have similar biochemistry, and perform analogous steps. Why, then, are replication proteins so divergent in amino acid sequence?

One possible explanation for the limited sequence similarity of replication proteins is that the cenancestor did not have a DNA genome but one based on RNA. Replication proteins would have thus evolved independently in the lineages leading to bacteria and archaea–eukaryotes after they split from a common ancestor. We think this unlikely for two reasons. First, despite the

general paucity of significant amino acid similarity between replication fork proteins of eubacteria and archaea–eukaryotes, some proteins are homologs, as discussed above.

Second, other components essential for replication, but not always situated at the replication fork, are also found in all three domains (Benner et al., 1989). These include proteins such as topoisomerases, gyrases, ribonucleases, and ribonucleotide reductases. In most cases, it is clear that analogous functions are performed by homologous proteins. Thus, multiple components involved in DNA replication, both at the replication fork and elsewhere, can be traced back to the cenancestor. We think it likely that the cenancestor was a DNA-based organism with a working DNA replication apparatus of some sort, but because of the lack of sequence similarity of bacterial and archaeal–eukaryotic replication proteins, we cannot confidently say what kind of DNA replication apparatus it was. The (not mutually exclusive) possibilities are as follows. (i) Our arguments notwithstanding, most bacterial and archaeal–eukaryotic replication proteins are homologous—they do descend from cenancestral proteins performing the same function—but have often been so radically changed in sequence as to be unrecognizable as homologs. (ii) The cenancestor contained both bacterial and archaeal–eukaryotic-type replication systems (perhaps one was for repair), and different components of these systems were lost in the bacterial and archaeal–eukaryotic lineage after their divergence. (iii) "New" (nonhomologous) proteins have been recruited into a replication function in one or the other lineages, replacing cenancestral components.

The Central Conundrum

Replication thus joins transcription and translation in compounding the central conundrum of cellular evolution, illustrated in Figure 1. The archaeal molecular components responsible for these fundamental cellular processes are more similar to their eukaryotic than their bacterial counterparts. We expect this in a quantitative sense if the structure of the universal tree shown in Figure 1 is correct—in fact, such quantitative similarity is the basis of the tree. But there are also qualitative differences between bacteria and archaea–eukaryotes—whole suites of proteins and protein complexes present in the one lineage and absent from the other. Often, the archaeal–eukaryotic system seems more complex (involves more components). And, for much of the latter half of this century we have thought of such complexity as part and parcel of the greater complexity in cellular ultrastructure (in particular, the presence of a cytoskeleton and endomembrane system) that distinguishes eukaryotes from all prokaryotes, archaea included.

Perhaps there is no conundrum—perhaps we were just wrong to think that way. But still, the uncoupling of molecular complexity in transcription, translation, and replication from complexity in cell structure is most peculiar. The eukaryotic cytoskeletal system seems to have arisen full blown at the origin of the eukaryotes. Although bacterial and archaeal FtsZ is a likely candidate for the prokaryote homolog of tubulin, amino acid alignments between these proteins are not very convincing. Russell Doolittle (Doolittle, 1995) has noted that the evolution of the eukaryotic cytoskeleton would require an abrupt change in the rate of sequence evolution for tubulins and their relatives that is of the order of "10- to 100-fold higher" than the observed rate of sequence evolution of tubulins and FtsZ proteins within the domains Bacteria, Archaea, and Eucarya. This rate of sequence evolution might also be invoked if DNA replication proteins of the three domains are in fact homologs and diverged from a common ancestral set of proteins.

Selected Reading

Belfort, M., and Weiner, A. (1997). Cell, this issue, *89*, 1003–1006.

Benner, S.A., Ellington, A.D., and Tauer, A. (1989). Proc. Natl. Acad. Sci. USA *86*, 7054–7058.

Bochkarev, A., Pfuetzner, R.A., Edwards, A.M., and Frappier, L. (1997). Nature *385*, 176–181.

Braithwaite, D.K., and Ito, J. (1993). Nucleic Acids Res. *21*, 787–802.

Budd, M.E., and Campbell, J.L. (1997). Mol. Cell. Biol. *17*, 2136–2142.

Bult, C.J., et al. (1996). Science *273*, 1066–1073.

Cullmann, G., Fien, K., Kobayashi, R., and Stillman, B. (1995). Mol. Cell. Biol. *15*, 4661–4671.

Dennis, P.P. (1997). Cell, this issue, *89*, 1007–1010.

Diffley, J.F.X. (1996). Genes Dev. *10*, 2819–2830.

Doolittle, R.F. (1986). Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences (Mill Valley, California: University Science Books).

Doolittle, R.F. (1995). Philos. Trans. R. Soc. (Lond.) B *349*, 235–240.

Doolittle, W.F. (1996). Proc. Natl. Acad. Sci. USA *93*, 8797–8799.

Forterre, P., and Elie, C. (1993). In The Biochemistry of Archaea (Archaebacteria), M. Kates, D.J. Kushner, and A.T. Matheson, eds. (Amsterdam: Elsevier), pp. 325–361.

Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J., Bowman, B., Manolson, M., Poole, R., Date, T., Oshima, T., et al. (1989). Proc. Natl. Acad. Sci. USA *86*, 6661–6665.

Iwabe, N., Kuma, K.-I., Hasegawa, M., Osawa, S., and Miyata, T. (1989). Proc. Natl. Acad. Sci. USA *86*, 9355–9359.

Kearsey, S.E., Maiorano, D., Holmes, E.C., and Todorov, I.T. (1996). Bioessays *18*, 183–190.

Kelman, Z., and O'Donnell, M. (1995). Nucleic Acids Res. *23*, 3613–3620.

Kornberg, A., and Baker, T.A. (1992). DNA Replication, 2nd. Ed. (New York: W. H. Freeman and Company).

Lieber, M.R. (1997). Bioessays *19*, 233–240.

Margolin, W., Wang, R., and Kumar, M. (1996). J. Bacteriol. *178*, 1320–1327.

Nolling, J., van Eeden, F.J., Eggen, R.I., and de Vos, W.M. (1992). Nucleic Acids Res. *20*, 6501–6507.

O'Donnell, M., Onrust, R., Dean, F.B., Chen, M., and Huwitz, J. (1993). Nucleic Acids Res. *21*, 1–3.

Philipova, D., Mullen, J.R., Maniar, H.S., Lu, J., Gu, C., and Brill, S.J. (1996). Genes Dev. *10*, 2222–2233.

Prasartkaew, S., Zijlstra, N.M., Wilairat, P., Prosper Overdulve, J., and de Vries, E. (1996). Nucleic Acids Res. *24*, 3934–3941.

Reeck, G.R., de Haën, C., Teller, D.C., Doolittle, R.F., Fitch, W.F., Dickerson, R.E., Chambon, P., McLachlan, A.D., Margoliash, E., Jukes, T.H., and Zuckerkandl, E. (1987). Cell *50*, 667.

Reeve, J.N., Sandman, K., and Daniels, C.J. (1997). Cell, this issue, *89*, 999–1002.

Stillman, B. (1994). Cell *78*, 725–728.