© 1999 Oxford University Press
Nucleic Acids Research, 1999, Vol. 27, No. 17 **3389–3401**

SURVEY AND SUMMARY

Did DNA replication evolve twice independently?

Detlef D. Leipe, L. Aravind[1] and Eugene V. Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, Bethesda, MD 20894, USA and [1]Department of Biology, Texas A&M University, College Station, TX 70843, USA

Received April 13, 1999; Revised and Accepted June 21, 1999

ABSTRACT

DNA replication is central to all extant cellular organisms. There are substantial functional similarities between the bacterial and the archaeal/eukaryotic replication machineries, including but not limited to defined origins, replication bidirectionality, RNA primers and leading and lagging strand synthesis. However, several core components of the bacterial replication machinery are unrelated or only distantly related to the functionally equivalent components of the archaeal/eukaryotic replication apparatus. This is in sharp contrast to the principal proteins involved in transcription and translation, which are highly conserved in all divisions of life. We performed detailed sequence comparisons of the proteins that fulfill indispensable functions in DNA replication and classified them into four main categories with respect to the conservation in bacteria and archaea/eukaryotes: (i) non-homologous, such as replicative polymerases and primases; (ii) containing homologous domains but apparently non-orthologous and conceivably independently recruited to function in replication, such as the principal replicative helicases or proofreading exonucleases; (iii) apparently orthologous but poorly conserved, such as the sliding clamp proteins or DNA ligases; (iv) orthologous and highly conserved, such as clamp-loader ATPases or 5'→3' exonucleases (FLAP nucleases). The universal conservation of some components of the DNA replication machinery and enzymes for DNA precursor biosynthesis but not the principal DNA polymerases suggests that the last common ancestor (LCA) of all modern cellular life forms possessed DNA but did not replicate it the way extant cells do. We propose that the LCA had a genetic system that contained both RNA and DNA, with the latter being produced by reverse transcription. Consequently, the modern-type system for double-stranded DNA replication likely evolved independently in the bacterial and archaeal/eukaryotic lineages.

INTRODUCTION

DNA replication is an essential, central feature of cellular life. There are many important functional parallels among all known cellular systems of DNA replication. These common features can be roughly summarized as follows: (i) replication is semi-conservative; (ii) replication always initiates at defined origins with the participation of an origin recognition system; (iii) replication fork movement is typically bidirectional; (iv) replication is continuous on the leading strand and discontinuous on the lagging strand; (v) RNA primers are needed to start DNA replication; (vi) nucleases, polymerases and ligases replace the RNA primers with DNA and seal the remaining nicks (1,2). It is therefore surprising that the protein sequences of several central components of the DNA replication machinery, above all the principal replicative polymerases, show very little or no sequence similarity between bacteria and archaea/eukaryotes (3,4). These observations suggest that some of the replication system components may not be homologs at all, whereas others, while homologous, are highly diverged. This is in stark contrast to the highly significant sequence similarity between the principal components of the transcription machinery, such as the DNA-dependent RNA polymerase (DdRp) subunits and a number of translation apparatus components.

The last 10 years have witnessed significant progress in our understanding of the relationships between proteins and domains involved in DNA replication. Significant sequence similarity between the polymerase-associated proofreading exonucleases of pro- and eukaryotes was noted in early studies (5). The recognition of homology between other replication proteins where sequence similarity was initially hard to detect has been made possible by structural comparisons. This was the case for the sliding clamp (6,7), the single-stranded (ss)DNA-binding proteins (8–10) and the 5'→3' (flap) endonucleases (11–14). No sequence similarity, however, has been detected between the principal replicative polymerases, namely the eubacterial family C (pol III) and the archaeal/eukaryotic family B polymerases, despite intense scrutiny at the sequence level (15–17) and despite the increasing availability of polymerase structures, including pol I from *Escherichia coli* (18) and *Thermus aquaticus* (13), HIV reverse transcriptase (19), T7 RNA polymerase (20) and a family B polymerase from phage RB69 (21). In the same vein, no sequence similarity could be found between the eubacterial and archaeal/eukaryotic primases (22).

*To whom correspondence should be addressed. Tel: +1 301 435 5913; Fax: +1 301 480 9241; Email: koonin@ncbi.nlm.nih.gov

Thus, the pattern of sequence conservation and divergence displayed by the replication proteins is fundamentally different from the pattern observed in the translation and transcription systems. It seems most likely that the core of the translation and transcription machinery was established in the last common ancestor (LCA) of all extant cells and subsequent evolution in different divisions of life did not involve dramatic alterations of the ancient molecular foundation. In contrast, major changes have occurred in the peripheral components, such as transcription regulators. Conversely, the core replication machinery, including the main replicative DNA polymerase, primase and the gap-filling polymerase, shows no detectable conservation. Several of the peripheral components, however, are clearly homologous or even orthologous. An obvious, though radical, explanation of the observed disparity is that the LCA did not have a DNA genome and its entire genetic system was RNA based. This hypothesis, however, does not account for the fact that several proteins involved in DNA replication as well as enzymes of deoxyribonucleotide biosynthesis and the recombination ATPase RecA are homologous in all extant organisms (23–26).

Edgell and Doolittle (4) delineated three distinct scenarios that could explain the existence of two versions of the replication machinery without invoking an RNA-only LCA. (i) The bacterial and archaeal/eukaryotic replicative systems have evolved from the LCA replication apparatus and the main replicative enzymes are actually homologs but, for some reason, have diverged rapidly and, in several cases, beyond recognition. (ii) The LCA possessed both a bacterial-type and an archaeal/eukaryotic-type DNA replication system (one of these could be responsible for repair) and the existence of two radically different systems in extant cells is due to differential gene loss in the bacterial and the archaeal/eukaryotic lineages. (iii) Either the bacterial or the archaeal/eukaryotic replication system is the direct descendant of the ancestral replication apparatus whereas the other version evolved by recruitment of non-homologous proteins accompanied by replacement of ancestral components.

To reach a clearer understanding of the origin(s) of the DNA replication system by comparative analysis of the sequences and structures of their components, additional, systematic effort in two directions seems to be necessary: (i) detecting subtle sequence and structural similarities that have escaped detection previously; (ii) solving the issue of orthologous relationships between replication components. The importance of the former aspect is underscored by the homologous relationship between the bacterial and eukaryotic sliding clamp proteins that was not originally recognized but became apparent when their structures had been determined (7,27). With the advent of more powerful methods for sequence analysis, such as PSI-BLAST (28), the similarity between the clamp proteins has become detectable at the sequence level. This suggests that systematic, careful comparisons of replication proteins might reveal additional subtle but evolutionarily and functionally important similarities. Such findings could shift the balance in our thinking about the evolution of DNA replication towards the common origin hypothesis, whereas the absence of detectable similarity in spite of a careful comparison might suggest independent origin for at least some of the components. It is critical for any meaningful evolutionary reconstruction to distinguish orthologs that likely evolved from an ancestral component of the replication machinery from homologous but not orthologous proteins that might have independently originated from proteins that had functions other than DNA replication.

With these considerations in mind, we attempted an exhaustive comparison of the sequences and structures of bacterial, archaeal and eukaryotic proteins known to be directly involved in DNA replication. We classified these proteins into orthologs, non-orthologous homologs and those components that appear to be completely unrelated. On the basis of this analysis, we propose a hypothesis that the LCA possessed a genetic system that involved both RNA and DNA, with the latter being produced by reverse transcription. Consequently, the modern-type system for double-stranded (ds)DNA replication might have evolved independently in the bacterial and archaeal/eukaryotic lineages.

DATABASES AND SEQUENCE ANALYSIS

For all sequence searches, the non-redundant database (NR) at the National Center for Biotechnology Information (NIH, Bethesda, MD) was used. The protein sequence similarity searches were performed using the gapped BLAST program and the PSI-BLAST program (28). The PSI-BLAST program constructs a position-dependent weight matrix from multiple alignments generated from the BLAST hits above a certain expectation value (e-value) and carries out iterative database searches using the information derived from this matrix (28,29). Normally, an e-value of 0.01 is considered an indication that a database hit is statistically significant after regions of low compositional complexity that tend to produce artifactually low e-values in database searches have been masked in the query sequence (29,30). Compositionally biased regions in protein sequences were masked prior to searches using the SEG program (31). The taxonomic breakdown of the database hits was produced using the Tax_Collector program of the SEALS package (32). The likely orthologs were identified on the basis of consistent inter-genomic best hits as described previously (33,34) and derived shared characters (synapomorphies) manifest at the level of distinct sequence motifs or features of domain architectures; the reasoning behind the assignment of orthologs is discussed below for each individual case.

EVOLUTIONARY RELATIONSHIPS BETWEEN BACTERIAL AND ARCHAEL/EUKARYOTIC DNA REPLICATION SYSTEMS

Table 1 lists the best database hits from archaea and eukaryotes for the principal bacterial proteins that are involved in DNA replication and DNA precursor synthesis in bacteria and in archaea/eukaryotes; analogous data for transcription machinery components are included as a control. Only a minority of the bacterial DNA replication machinery components show significant similarity to archaeal/eukaryotic homologs. Some of the strongest hits from bacteria to eukaryotes, such as those to the human NAD-dependent DNA ligase and the pol I homolog from *Drosophila*, are readily explained by horizontal gene transfer, most likely from organelles (see also 35). Additional cases of likely horizontal transfer, apparently from eukaryotes or archaea to bacteria, are seen in a reciprocal analysis of eukaryotic replication machinery components. These include the B family DNA polymerases, which are ubiquitous in

Table 1. Bacterial, archaeal and eukaryotic homologs of *E.coli* DNA replication machinery components[a]

E. coli protein	Best hit in Bacteria (-Proteobacteria), e-value[b]	Best hit in Archaea, e-value[b]	Best hit in eukaryotes (-organelles), e-value[b]	Function/Comment
Replication				
DnaA	**118794_Bs; 6e-94**	-	-	**ATPase required for the initiation of replication**
DnaC	**2983431_Aa; 1e-04**	-	-	**Accessory ATPase involved in initiation**
PriA	**3183549_Bs; 1e-114**	-	-	**Primosomal DNA helicase**
DnaG	**464463_Ca; 9e-71**	-	-	**Primase**
DnaB	**585057_Bs; 1e-108**	-	-	**Main replicative helicase**
UvrD	3024353_Bst; 1e-152	**2621541_Mth; 4e-50**	1723281_Sp; 6e-68	Accessory replicative helicase
Rep	3024353; 1e-121	**2621541;1e-45**	1723281_Sp; 2e-65	Accessory replicative helicase
DnaE_PHP	**3913509_Bb; 2e-19**	-	-	**Replicative polymerase, predicted phosphatase domain**
DnaE_pol	**3913509_Bb; 0.0**	-	-	**Replicative polymerase, polymerization domain**
DnaQ	3322942_Tp; 2e-17	**2649627_Af; 7e-10**	-	3'-5'-proofreading exonuclease associated with the replicative polymerase
DnaN	**3328470_Ct; 4e-42**	-	-	**Sliding clamp subunit of DNA polymerase**
DnaZX	580855_Bs; 3e-67	2621290; 8e-18	4220511; 1e-39	Clamp loader ATPase
HolB	**2105050_Mt; 4e-11**	-	-	**Accessory subunit of the clamp loader**
PolA_5'exo	416913_Bc; 1e-48	-	**3845116_Pf; 3e-09**	Gap-filling DNA polymerase, 5'-3' exonuclease domain. Primer removal
PolA_3'exo	1913934-Cau; 3e-22	-	-	Gap-filling DNA polymerase, 3'-5' proofreading exonuclease domain
PolA_pol	4090935_Rsp; 1e-110	-	**4105277_Hs; 6e-58**	Gap-filling polymerase, polymerization domain
Rnh	581811_Tt; 4e-44	-	2677845_Dm; 4e-19	RNAase HI, primer removal

eukaryotes and archaea but so far present only in the γ-proteobacterial lineage, and ATP-dependent DNA ligases, which show a sporadic presence in certain bacteria (data not shown).

These cases of apparent horizontal gene transfer apart, the striking contrast between the replication and transcription systems, in terms of conservation of the respective components (or lack thereof), in bacteria and archaea/eukaryotes is obvious (Table 1). Although both the replication system and the transcription system include proteins that are highly conserved between bacteria and archaea/eukaryotes, along with ones that show little or no similarity, the breakdown of these systems into conserved and distinct components goes along very different lines. In the transcription machinery, the principal subunits of the DdRp show high levels of conservation, whereas accessory polymerase subunits and transcription factors are poorly conserved or show no detectable similarity at all. Amongst the replicative proteins, the situation is inverted; the DNA polymerases and primases are not detectably similar and only some of the accessory subunits, such as clamp-loading ATPases, enzymes that participate in replication but are not components of the replication fork, such as topoisomerase I, and at least some DNA precursor biosynthesis enzymes are highly conserved (Table 1).

To solve the central conundrum in the evolution of replication—common versus independent origins of the bacterial and archaeal/eukaryotic systems—it is not enough to show that components of the DNA replication machinery are homologous or non-homologous. Replication of dsDNA poses a number of

Table 1. *Continued.*

RnhB	2633978_Bs; 2e-40	3551209_Pk; 5e-12	3879811_Ce; 3e-07	RNAase HII, primer removal
Lig	26329761_Bs; 1e-175	-	**1770452_Hs; 3e-96**	DNA ligase
Ssb	**586039_Bs; 1e-13**	-	-	SsDNA-binding protein
TopA	520753_Bs; 1e-142	592234_Mj; 4e-62	2827516_At; 2e-82	Topoisomerase I, supercoiling relaxation
GyrA	80350_Bs; 0.0	**2650163_Af; 0.0**	3172113_En; 5e-18	DNA gyrase (topoisomeraseII), subunit A
GyrB	80348_Bs; 0.0	**2650095_Af; 0.0**	2129576_At; 3e-38	DNA gyrase (topoisomeraseII), subunit A
RecA	154653_Tf; 0.0	2665476_Mm; 1e-07	2058711_Bm; 2e-11	Recombinase, ATP-dependent strand annealing
DNA precursor biosynthesis[c]				
NrdE	3261509_Mt; 0.0	2648891_Af; 3e-34	200765_Mmu; 2e-46	Ribonucleotide reductase 2, α-subunit
NrdF	421244_Mt; 1e-165	-	1044912_Nt; 3e-07	Ribonucleotide reductase 2, β-subunit
NrdD	4098081_Ll; 0.0	2622659_Mth; 8e-34	-	Anaerobic ribonucleotide reductase
ThyA	143741_Bs; 1e-104	026868_Mth; 2e-04	1361867_Mmu; 8e-78	Thymidylate synthase
Transcription				
RpoB	3328732_Ct; 0.0	3122768_Mth; 1e-33	3603015_Gt; 0.0	DNA-directed RNA polymerase β-subunit
RpoC	3328731_Ct; 0.0	3257973_Ph; 3e-50	4092885_Nl; 1e-38	DNA-directed RNA polymerase β'-subunit
RpoA	133395_Bs; 7e-70	3258066_Ph; 0.004	-	DNA-directed RNA polymerase α-subunit; eukaryotic orthologs detectable in iterative searches
RpoD	**2258087_Sm; 3e-88**	-	-	**DNA-directed RNA polymerase σ-subunit**
NusA	2634032_Bs; 2e-68	139955_Hh; 6e-05	-	Transcription antitermination factor
NusB	**1709418_Bs; 3e-15**	-	-	**Transcription termination factor**
NusG	**548391_Bs; 5e-41**	-	-	**Transcription antitermination factor**
GreA	**3183527_Bs; 3e-25**	-	-	**Transcription elongation factor**
Rho	3322527_Tp; 1e-141	2605826_Mb; 4e-05	-	Transcription termination factor, RNA helicase. The archaeal homologs are H^+-ATPase subunits; the similarity to eukaryotic vacuolar ATPase subunits is detectable in iterative searches

similar problems in any system and it would not be unexpected if independently evolving solutions were similar, given that ancient protein superfamilies, such as the P-loop ATPases, were already available for recruitment in the LCA. Thus the goal of comparative analysis of the replication systems is to distinguish, as best we can, between those components that appear to be orthologous and thus should have descended from an LCA protein that had the same function and those for which,

Table 1. *Continued.*

[a]Analogous data for selected enzymes of DNA precursor biosynthesis and principal proteins involved in transcription are included for comparison. Data for accessory proteins that are not highly conserved among bacteria are not shown.
[b]e – n = 10^{-n}; e-values more significant than 1e – 180 are given as 0; a dash is shown if no significant BLAST hit has been found for the given lineage (e-value cut-off 0.1). Yellow shading shows proteins with significant hits only in bacteria and pink shading denotes likely horizontal gene transfers (see text). For each lineage-specific best hit, the Gene Identification number and the species name abbreviation are given. Aa, *Aquifex aeolicus*; Af, *Archaeoglobus fulgidus*; At, *Arabidopsis thaliana*; Bb, *Borrelia burgdorferi*, Bm, *Bombyx mori*; Bs, *Bacillus subtilis*; Bsp, *Bacillus* sp., Bst, *Bacillus stearothermophilus*; Ca, *Clostridium acetobutylicum*; Cau, *Chloroflexus auranticus*; Ce, *Caenorhabditis elegans*; Ct, *Chlamydia trachomatis*; Dm, *Drosophila melanogaster*; Gt, *Guillardia theta*; Hh, *Halobacterium halobium*; Ll, *Lactococcus lactis*; Mb, *Methanosarcina barkeri*; Mm, *Methanococcus maripaludis*; Mmu, *Mus musculus*; Mt, *Mycobacterium tuberculosis*; Mth, *Methanobacterium thermoautotrophicum*; Nl, *Nosema locustae*; Nt, *Nicotiana tabaccum*; Ph, *Pyrococcus horikoshii*; Pk, *Pyrococcus kodakoraensis*; Rsp, *Rhodotermus* sp.; Sm, *Streptococcus mutans*; Tp, *Treponema pallidum*; Tt, *Thermus thermophilus.*
[c]The complex evolution patterns of enzymes of DNA precursor biosynthesis are beyond the scope of this work; we present the data for only two types of key enzymes, to emphasize their conservation in bacteria, archaea and eukaryotes.

whether they are homologous or not, independent origin is more likely. Proving independent origin is hard, if at all possible. The case, however, is strongly supported if, for example, an archaeal/eukaryotic protein with a central role in replication is most closely related not to its bacterial functional counterpart but to a protein family that performs functions outside replication.

The lack of detectable sequence similarity does not automatically mean that the respective proteins are not homologs; there are examples of very subtle relationships between bacterial and archaeal/eukaryotic proteins that nevertheless appear to indicate homology or even orthology (see for example 36). Conversely, even highly significant sequence similarity, such as that observed between the clamp-loader ATPases, is not necessarily a guarantee of orthology.

With these considerations in mind, we performed a more detailed, case-by-case analysis of the bacterial, archaeal and eukaryotic proteins involved in DNA replication. Figure 1 summarizes the domain arrangements seen in the protein components of the bacterial, archaeal and eukaryotic DNA replication machineries and the relationships between them. In Table 2, replicative proteins are classified into four principal categories that are discussed below.

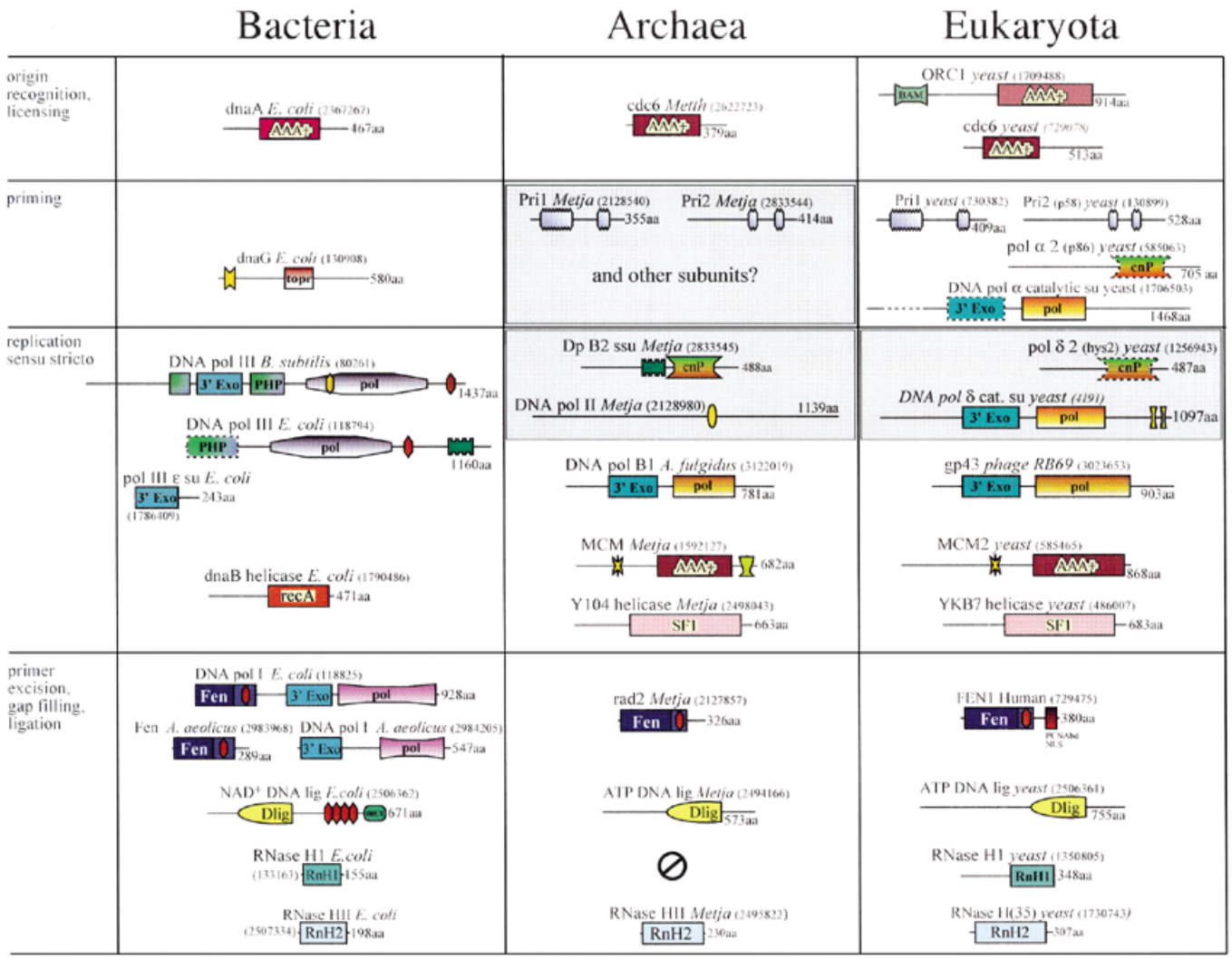
Unrelated components in the bacterial and archaeal/ eukaryotic DNA replication machineries

This category consists of only four domains but, strikingly, these include all three functional types of DNA polymerases required for replication, namely the DNA polymerase involved in elongation, the primase that is responsible for primer synthesis and hence the initiation of DNA replication, and the DNA polymerase involved in gap-filling upon primer removal. Not only database searches but also direct comparisons fail to show any sequence similarity between the nucleotide polymerization domain of bacterial DNA polymerase III (pol III) α-subunit and the functionally analogous domain of the archaeal and eukaryotic family B DNA polymerases (or any other proteins). The same is true of the second archaeal DNA polymerase (pol IV), whose large subunit, with the exception of a Zn-ribbon domain, appears to be unrelated to either bacterial or eukaryotic polymerases (37,38). The 3-dimensional structures of pol III and pol IV have not been determined and therefore it cannot be ruled out that they have the 'palm-and-fingers' structure similar to that seen in other DNA polymerases, including the bacteriophage RB69 polymerase (21), which represents the archaeal/eukaryotic family B. However, counterparts to the conserved motifs that appear to be shared by the eukaryotic and archaeal DNA polymerases, reverse transcriptases and RNA-dependent RNA polymerases (RdRps) (16) are not detectable in pol III and pol IV. This makes a specific evolutionary affinity between the bacterial and archaeal/eukaryotic DNA polymerase subunits involved in chain elongation during DNA replication most unlikely.

Both types of replicative DNA polymerases possess two additional enzymatic domains that also may function as separate subunits, namely a 3'→5' exonuclease and a predicted phosphoesterase (Fig. 1). The exonuclease domains are related but may not be orthologous, as discussed below. In contrast, the phosphoesterase domains/subunits belong to two distinct enzyme superfamilies, namely the PHP superfamily in bacteria and the calcineurin-type superfamily of metal-dependent phosphoesterases in archaea and eukaryotes, which show no indication of a homologous relationship (39).

DNA primases present a case where an independent origin of the bacterial and archaeal/eukaryotic enzymes appears to be supported by positive evidence as well as a lack of detectable sequence similarity. The catalytic domain of bacterial primases shows a subtle but statistically significant sequence similarity to the DNA-nicking-rejoining domains of type I, type II and type VI topoisomerases and a distinct group of nucleases; all these proteins are predicted to contain the conserved Toprim domain (22). Despite a careful search, we were unable to detect any similarity to the Toprim domain in the sequences of eukaryotic primases. The fact that bacterial primases show an apparent structural and evolutionary relationship not with their archaeal/eukaryotic functional counterparts but with enzymes that have significantly different, even if mechanistically related, functions seems to effectively rule out an origin of the two types of extant primases from an ancestral primase.

Finally, the bacterial gap-filling DNA polymerase (pol I) appears to be unrelated (or, at best, extremely distantly



related), with the exception of the 3'→5' exonuclease domain, to any other DNA polymerases, whereas eukaryotes utilize family B DNA polymerases for both elongation and gap-filling (Fig. 1 and Table 2).

Homologous but not orthologous components of the DNA replication apparatus in bacteria and archaea/eukaryotes

Several important components of the DNA replication machinery in bacteria and archaea/eukaryotes, while homologous, are strong candidates for independent recruitment for a role in replication. The example of the principal replicative helicases is the most straightforward one. All helicases appear to be ultimately homologous as members of the P-loop NTPase fold (40–42). This generic relationship apart, however, the bacterial replicative helicase DnaB and the helicases involved in eukaryotic replication, such as the DNA polymerase α-associated helicase A from yeast (ORF YKL017c) (43), belong to different divisions of the P-loop NTPase fold. Yeast helicase A belongs to helicase superfamily I, which includes a variety of DNA and RNA helicases, such as, for example, bacterial UvrD, that are involved in repair functions and may also perform accessory roles in replication. Some of the highly conserved eukaryotic homologs of helicase A are RNA helicases, such as the NAM7/UPF1 proteins from fungi and animals, that are required for the processing of nonsense mRNAs (44,45), and yeast SEN1, that is involved in the endonucleolytic cleavage of introns from precursor tRNAs (46). Another group of highly conserved archaeal and eukaryotic DNA helicases involved in replication, the MCM proteins, belongs to the AAA+ superfamily of P-loop NTPases (42,47). In addition to the MCM helicases and the bacterial helicase RuvB, involved in repair, this superfamily includes a variety of ATPases with broadly defined chaperone-like functions, e.g. subunits of ATP-dependent proteases. In contrast, DnaB is a member of a distinct family that is specifically related to the RecA family, to the exclusion of other groups of ATPases (Table 2; D.D.Leipe, L.Aravind and E.V.Koonin,

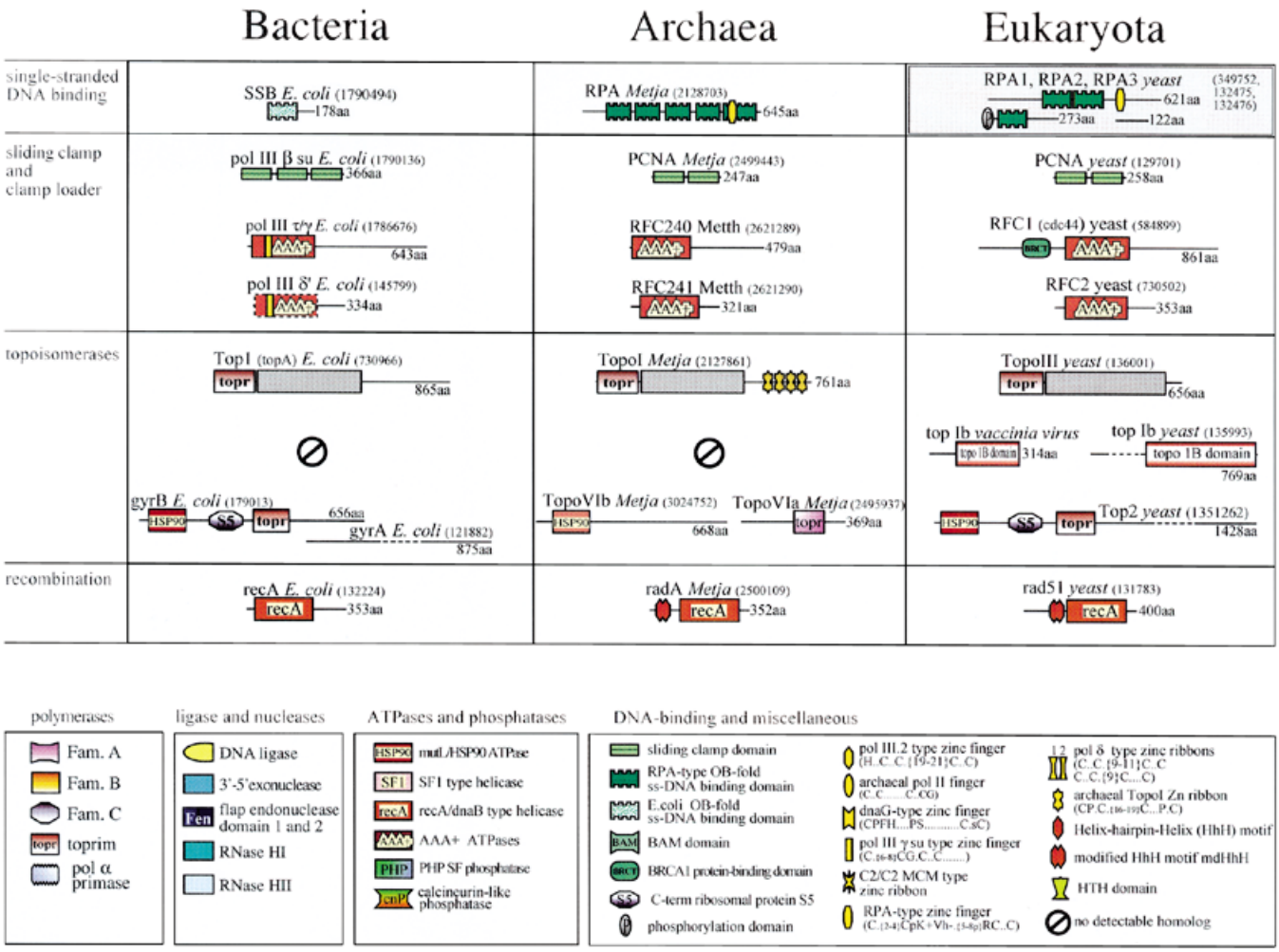


Figure 1. (Opposite and above) Domain organization of the principal proteins involved in DNA replication in bacteria and archaea/eukaryotes. Proteins are represented as horizontal lines and regions of sequence similarity between eubacterial and archaeal/eukaryotic proteins are indicated by rectangles or other geometric shapes. Domains (proteins) considered orthologous are shown with the same label and color (e.g. Fen nuclease domain). Each sequence is identified by gene or protein name, organism name and the GenBank identifier (in parentheses). Domains that contain substitutions in (predicted) catalytic residues and accordingly are perceived to be inactive are shown framed by broken lines. The identity of the archaeal/eukaryotic replication fork helicase is uncertain and the figure depicts the two likely candidates, MCM and YKB7. Dlig, DNA ligase domain; pol, DNA polymerase catalytic domain; S5, domain similar to C-terminal domain of ribosomal protein S5 (63); SF1, superfamily 1 helicase; HSP90/mutL, domain found in the mutL ATPase (64); HhH, helix–hairpin–helix DNA-binding motif (65); BRCT, BRCA1 C-terminus (66); Fen, flap nuclease domains (11,67); pol III τ/γ, pol III δ', RFC1, RFC2, clamp-loader subunits; Pri, primase; Toprim, topoisomerase–primase catalytic domain (22); 3' Exo, 3'→5' proofreading exonuclease domain/protein. Sequence conservation in the zinc-binding domain in pol III is compatible with the formation of two distinct finger structures, a 29 residue Cys4 finger or a 32 residue His1/Cys3 finger (68). Organisms are designated as follows: *A. aeolicus*, *Aquifex aeolicus*; *A. fulgidus*, *Archaeoglobus fulgidus*; *B. subtilis*, *Bacillus subtilis*; *E. coli*, *Escherichia coli*; *Metja*, *Methanococcus jannaschii*; *Metth*, *Methanobacterium thermoautotrophicum*; Human, *Homo sapiens*; yeast, *Saccharomyces cerevisiae*; *phage RB69*; *vaccinia virus*.

unpublished observations). Thus, the principal replicative helicase seems to be an irrefutable case of independent drawing of enzymes from the pool of P-loop ATPases for a central function in DNA replication.

The case of the origin recognition and licensing ATPases is more complicated in that the protein that performs this function in bacteria (DnaA), its functional analogs in eukaryotes (the origin recognition complex subunits, e.g. ORC1) and their archaeal homologs all belong to the AAA+ superfamily of P-loop ATPases (42). Within this superfamily, however, DnaA does not cluster with its functional counterparts from eukaryotes or archaea, suggesting that there is no orthologous relationship between the bacterial and archaeal/eukaryotic origin recognition ATPases (Table 2).

Table 2. Relationships between the principal components of the DNA replication machinery in bacteria and archaea/eukaryotes

Function	Archaeal/ Eukaryotic protein	Bacterial protein (*E. coli*)	Best archaeal or eukaryotic (-organelles) hits for the *E. coli* protein; e-value (PSI-BLAST iteration)[a]	Comment
Apparently unrelated components				
Main replicative polymerase, polymerization domain	B family polymerases	PolIII (DnaE_pol)	None	
Main replicative polymerase, predicted phosphatase domain (subunit)	Calcineurin-type superfamily phosphatase	Predicted PHP superfamily phosphatase (DnaE_PHP)	2496191_Mj; 2e-04 (1)	The archaeal homologs are uncharacterized proteins predicted to possess phosphatase activity. Yeast histidinol phosphatase is a member of the PHP superfamily; the similarity between this protein and the PHP domain of PolIII is detectable in searches started with other members of the superfamily (39).
Gap-filling DNA polymerase, polymerization domain	DNA polymerase ε or τ	DNA polymerase I (PolA_pol)	See Table 1	There is no archaeal homologs of bacterial PolI. The eukaryotic protein containing the PolI domain fused with a helicase is involved in repair rather than replication (35); the gene coding for this enzyme is a likely horizontal transfer from bacteria (organelles)
DNA primase	DNA polymerase α-associated, 2-subunit primase	DnaG-type primase	3258129_Ph; 1e-04(1)	Given the presence of the orthologs of the two eukaryotic primase subunits, archaeal DnaG homologs are implicated in repair (22). The genes for these proteins might have been horizontally transferred from bacteria. The closest eukaryotic similarity is the Toprim domain of topoisomerases which is not detectable with bacterial primases as starting points (22).
Distantly related, non-orthologous components				
ATPase involved in initiation	ORC1; MCM proteins?	DnaA	2492505_Mj; 3e-05 (2)	The closest homologs of DnaA detected by PSI-BLAST analysis are eukaryotic CDC48 ATPases that are associated with endoplasmic reticulum and are involved in cell cycle control, and their archaeal ortholog
Replicative helicase	Helicase A (YKB7_YEAST), MCM proteins?	DnaB	1142660_Pc (DnaB ortholog); 6e-07 (1); 3107925_Hs (RecA/RadA	The function of eukaryotic DnaB orthologs is not known; this gene is missing in yeast and accordingly, is not an essential part of the replication system. Origin by

An even more complex relationship is seen between the 3'→5' proofreading exonucleases of bacterial and archaeal/eukaryotic replicative polymerases. In bacteria they exist either separately as the ε subunit of pol III or are inserted into the PHP domain of one of the multiple α-subunits of pol III in the Gram-positive lineage and *Thermotoga* (Fig. 1). In the archaea and eukaryotes, the 3'→5' exonuclease is always fused to the DNA polymerase catalytic domain. Both bacterial and archaeal/eukaryotic proofreading exonucleases belong to the large superfamily of 3'→5' exonucleases that includes not only DNases but also a variety of RNases (48). Phylogenetic tree analyses do not show enough resolution to meaningfully address the issue of the monophyly of the proofreading exonucleases to the exclusion of other nucleases in this superfamily (data not shown). The sequence similarity between the exonuclease domains of bacterial pol III and archaeal/eukaryotic polymerases is low (two to four iterations of PSI-BLAST are required to detect it). Bacterial pol III proofreading enzymes show the greatest similarity to a group of eukaryotic poly(A)-processing enzymes. The 3'→5' exonuclease domains fused to

Table 2. *Continued.*

			family ATPase); 8e-07 (1)	horizontal transfer from bacteria (organelles) is likely. The relationship with the RecA family is consistently detectable and suggests an origin of DnaB from RecA (D. D. Leipe, L. Aravind and E. V. Koonin, unpublished observations; also see text).
ssDNA-binding protein	RPA protein (multiple OB-fold domains)	Ssb (OB-fold domain)	None	The similarity is apparent at the structural level but is not readily demonstrable at the sequence level (see text)
Main replicative DNA polymerase, 3'-5'-exonuclease domain	3'-5'-exonuclease domain of Family B DNA polymerases	3'-5'-exo domain of polIII	4007761_Sp; 4e-07(1)	The closest eukaryotic homolog appears to be an Rnase involved in splicing. However, an orthologous relationship between exonuclease domains of polymerases cannot be ruled out despite low similarity (see text).
Gap-filling DNA polymerase, 3'-5'-exo domain	3'-5'-exo domain of Family B DNA polymerases	3'-5'-exo domain of polI	1723221_Sp; 3e-07 (2)	The closest eukaryotic homolog appears to be an RNase involved in splicing. However, an orthologous relationship between exonuclease domains of polymerases cannot be ruled out despite low similarity (see text).
Apparently orthologous but distantly related components				
Sliding clamp subunit of DNA polymerase	Proliferating cell nuclear antigen (PCNA) and its archaeal orthologs	PolIII β-subunit (DnaN)	2499443_Mj; 1e-04(1)	
DNA ligase	ATP-dependent ligase	NAD-dependent ligase	None (see Table 1 for likely horizontal transfer)	The similarity between NAD-dependent and ATP-dependent ligases is detectable in PSI-BLAST searches started with the latter (36). The C-terminal BRCT domain that is conserved in a variety of eukaryotic proteins (66) was not used for the search.
5'-3' exonuclease (flap nuclease)	Flap nuclease (FEN1, Rad2)	5'-3'-exo domain of PolI	1490870_Xl; 1e-06 (1); see also Table1	
Orthologous, highly conserved components				
Clamp-loader ATPase	Replication factor C	PolIII ZX-subunit	See Table1	
Topoisomerase I/III	Topoisomerase I/III	Topoisomerase I (swivelase)	See Table1	
RNAase HII			See Table1	
Recombinase	RadA	RecA	See Table1	

[a]Only hits appearing in iterative but not in single-pass searches are included; the highly significant hits seen in single-pass searches are given in Table 1. The other designations are as in Table 1. Hs, *Homo sapiens*; Mj, *Methanococcus jannaschii*; Pc, *Plasmodium chabadii*; Ph, *Pyrococcus horikoshii*; Sp, *Schizosaccharomyces pombe*; Xl, *Xenopus laevis*.

bacterial pol I and to helicases, such as the vertebrate Werner syndrome protein, are also significantly similar to this group, which suggests that these domains were recruited for different functions on multiple occasions. Given the high level of divergence and the abundance of RNases in the 3'→5' exonuclease superfamily, it is not certain whether the extant proofreading nucleases are all descendents of an ancestral proofreading enzyme or have been independently recruited for this task from the general pool of exonucleases.

The ssDNA-binding proteins represent another case of homologous domains that apparently have been independently recruited to perform a similar function in the archaeal/eukaryotic

and bacterial lineages. Both bacterial and archaeal/eukaryotic ssDNA-binding proteins contain the ancient, widespread nucleic acid-binding domains of the OB-fold (49). A detailed sequence comparison showed that the eukaryotic ssDNA-binding protein that contains three OB-fold domains and its archaeal counterpart containing five OB-fold domains (the RPA proteins) are most closely related to the subclass of OB-folds typified by those in the lysyl- and aspartyl-tRNA synthetases (L.Aravind, unpublished observations). Similar OB-folds are also found in bacterial pol III α-subunits, the small subunit of the archaeal DNA polymerases (N-terminal to the phosphoesterase domain) and some bacterial and archaeal nucleases. Thus the archaeal/eukaryotic ssDNA-binding proteins belong to a distinct family of OB-folds that includes both RNA- and DNA-binding members. In contrast, sequence comparisons show that bacterial ssDNA-binding proteins form a separate family of OB-folds with distinct structural features, such as unusually long β-strands (9,50).

Orthologous components of the bacterial and archaeal/eukaryotic replication machineries

A considerable subset of the proteins that comprise the replication machinery appears to be represented by orthologs in all extant organisms. In only two cases, however, namely those of RNase HII and topoisomerase IA, do these protein show obvious, high conservation at the sequence level (Table 1).

The bacterial and archaeal/eukaryotic clamp-loader ATPases show a moderate but statistically significant similarity to each other (Table 1). There are, however, considerable differences in the domain architectures of the bacterial and eukaryotic clamp-loaders, such as the presence of BRCT domains in eukaryotic but not bacterial clamp-loaders and, conversely, the presence of a zinc-finger in bacterial but not eukaryotic ones (Fig. 1). Nevertheless, the presence of unique sequence signatures, such as the SRC motif (42,51), suggests that the ATPase domains of the clamp-loaders are orthologous.

Other proteins and domains, namely archaeal/eukaryotic FEN1/RAD2 nucleases and bacterial 5'→3' exonuclease domains of polymerase I, the replication sliding clamps (PCNA) and DNA ligases (the NAD-dependent ligase in bacteria and the ATP-dependent ligase in eukaryotes), show very low sequence conservation but, nevertheless, appear to be orthologs (Table 2 and Fig. 1). Until recently, the homologous relationships between these components of the replication machinery remained undetected. However, detailed sequence comparisons as well as structural superposition for the sliding clamps and the ligases (36,52; see also above) indicated that in each of these cases, the bacterial and archaeal/eukaryotic proteins are homologous. Moreover, apparent horizontal gene transfers apart, the bacterial proteins in each of these cases are more similar to their functional counterparts from archaea/eukaryotes than to any other archaeal or eukaryotic proteins (Table 2). These observations suggest that orthologous relationships exist for each of these proteins, in spite of the high level of divergence.

Finally, some replication proteins, such as RNase H1 and topoisomerase II, are highly conserved in bacteria and eukaryotes but are missing from the Archaea. This distribution might be indicative of a horizontal transfer from bacteria to eukaryotes, although it cannot be ruled out that these proteins were present in the LCA and have been lost in the archaeal lineage. Furthermore, archaeal topoisomerase VI appears to be orthologous to eukaryotic proteins involved in recombination (e.g. yeast Spo11) but is only distantly related to bacterial and eukaryotic topoisomerase II (53). This suggests that the lack of a distinct archaeal topoisomerase II ortholog might be alternatively explained by extreme divergence.

Hypothesis: a mixed, RNA/DNA genetic system in the LCA

As discussed above, the DNA replication machinery in bacteria, compared to that of archaea/eukaryotes, is built from a patchwork of orthologous (but sometimes highly diverged) proteins, proteins that are homologous but apparently have been independently recruited for replication and a core of polymerases that seem to be unrelated (Table 2 and Fig. 1).

How can this mixture of ancestral and independently acquired features of the DNA replication systems be accounted for? Three principal models can be envisioned for the replication of the genome of the LCA. (i) The LCA had an RNA genome that was replicated by RdRp. (ii) The LCA already had a DNA genome, like modern-day cells, that was replicated by DNA-directed DNA polymerases (DdDp). (iii) The genome of the LCA had an RNA component and a DNA component, with the DNA being transcribed into RNA and RNA being reverse transcribed into DNA. Given the orthology and high conservation of the core components of the eubacterial and archaeal/eukaryotic transcription machinery, as well as the orthologous relationships between at least some enzymes of DNA precursor biosynthesis, several components of the replication machinery itself and the RecA/RadA recombinase, the first possibility seems unrealistic. The LCA must have been able to synthesize and make use of DNA. The second model must somehow explain the lack of orthology and, in several cases, any detectable homologous relationship whatsoever between key components of the DNA replication apparatus in bacteria compared to archaea/eukaryotes. As already mentioned, such explanations would involve one or more of the three main themes: (i) the principal components of the DNA replication are in fact orthologous in all forms of life but have diverged beyond recognition; (ii) there has been non-orthologous displacement of some but not other components of the DNA replication machinery in one of the divisions of life (e.g. bacteria); (iii) the LCA possessed two (partially) independent DNA replication systems that have been eliminated in a lineage-specific fashion during subsequent evolution.

The complexity of the eukaryotic chromatin in the form of linear chromosomes, larger genome size and higher order packaging does impose new problems on any DNA handling system (54). Such changes are visible in the basic repair enzymes (35) and transcription machinery of the eukaryotes and, in principle, might account for the rapid divergence of the replication systems. However, archaea have single circular chromosomes and genome size in the same range as bacteria but their replication machinery is orthologous to the eukaryotic one (with some important distinctions, such as the presence of a unique DNA polymerase) and dissimilar from the bacterial one, as discussed above. Thus the distinction between the bacterial and the archaeal/eukaryotic replication systems does not seem to correlate with the major changes in chromatin structure and genome organization which separate eukaryotes from both bacteria and archaea. The advent of the eukaryotic chromatin organization is associated with the recruitment of additional subunits to the replication complexes but not with

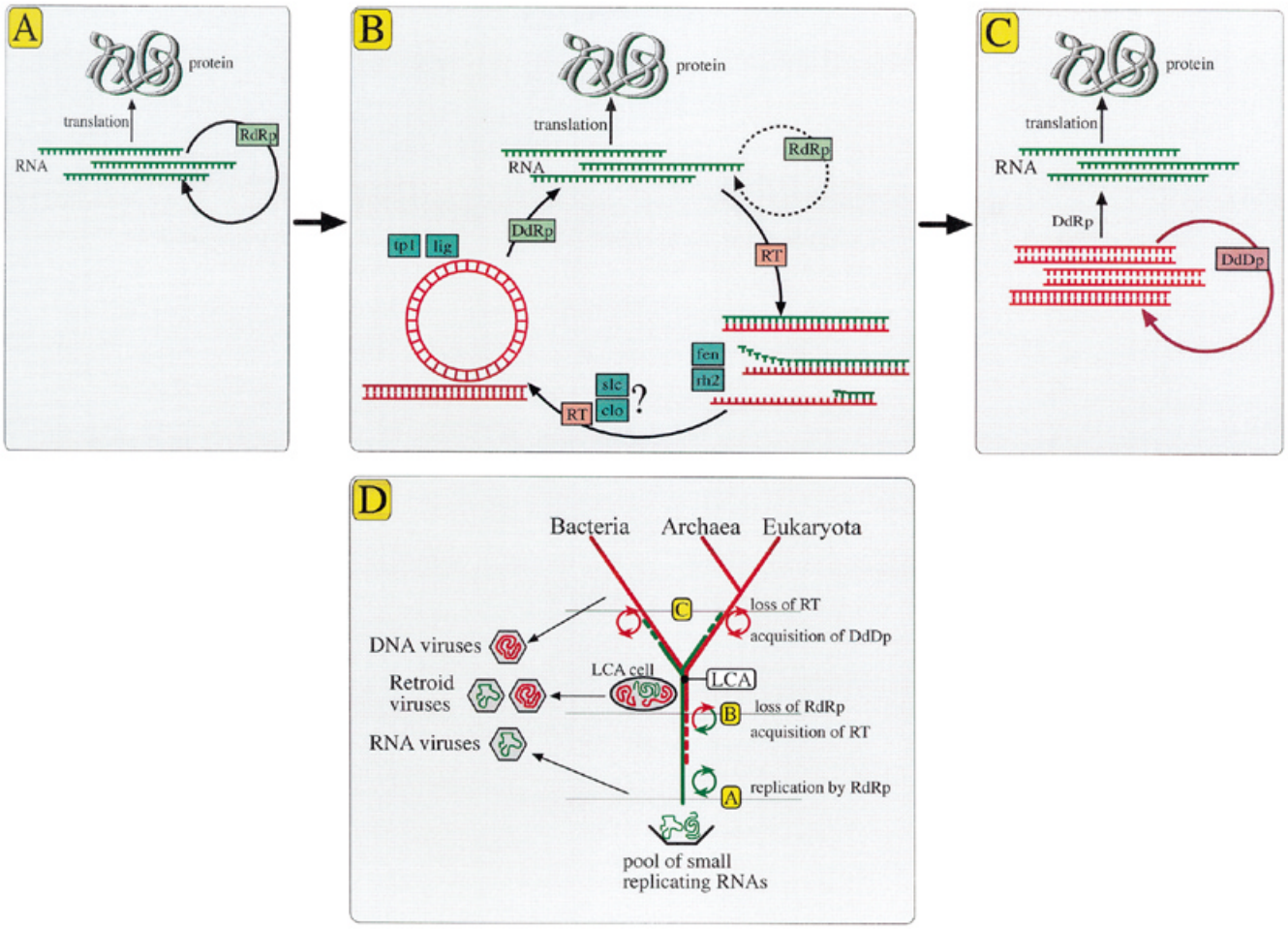


Figure 2. A hypothetical scenario for the evolution of the genetic system. (**A**) Ancient RNA-based system. (**B**) Mixed RNA/DNA system postulated for the LCA. (**C**) Modern-type dsDNA-based system. (**D**) A scheme of transition between the three postulated main stages in the evolution of replication. Early cells might have had an RNA genome that was replicated by RdRps (A). Conceivably, at an early stage of DNA usage, only ssDNA was generated from the RNA genome and functioned in the form of a RNA–DNA hybrid, while cells were still missing the capability to synthesize dsDNA. This hypothetical stage is not pictured here. The genome of the LCA has an RNA and a DNA component where DNA is transcribed into RNA and RNA is reverse transcribed into first ssDNA, then dsDNA (B). Circular DNA molecules, if present at this stage, could have necessitated the involvement of DNA ligases and topoisomerase. Modern cells replicate the DNA genome with DdDps (C). DNA is in red, RNA in green. Protein names are enclosed in rectangles: green, RNA polymerases; red, DNA polymerases; blue-gray, accessory proteins. clo, clamp-loader; DdDp, DNA-dependent DNA polymerase; RdDp, RNA-dependent DNA polymerase; fen, flap endonuclease; LCA, last common ancestor of Bacteria and Archaea/Eukaryota; lig, DNA ligase; RdRp, RNA-dependent RNA polymerase; rh2, RNase HII; RT, reverse transcriptase; scl, sliding clamp; tp1, type 1 topoisomerase. The yellow letters A–C in (D) refer to the stage of replication evolution depicted in the panel with the corresponding letter.

dramatic changes to the core components. This makes a major acceleration of evolution a highly unlikely explanation for the disparity between the replication systems of bacteria and archaea/eukaryotes.

Non-orthologous gene displacement, i.e. recruitment of genes from outside the replication machinery, offers an alternative way to account for the lack of sequence similarity between replication machinery proteins. For some of the replication proteins, a possible source for such recruitment exists, e.g. topoisomerases for bacterial-type DNA primases or AAA+ ATPases with chaperone functions for ATPases involved in replication (DnaA or ORC1). It is hard to imagine, however, what could be the selective advantage of the displacement of key components of the replication apparatus, particularly if such displacements were to occur one at a time. Simultaneous displacement of multiple components, in contrast, would effectively amount to a takeover by an independently evolved replication system which would mean two origins rather than one for the DNA replication machinery.

The third option, namely the differential loss of one of the two DNA replication systems inherited from the LCA (one of them originally responsible for repair), is perhaps most difficult to refute. However, in addition to being based on the unlikely assumption that the replication system of the LCA was considerably more complex than modern ones, this hypothesis also runs into problems with non-orthologous displacement

mentioned above, in this case with regard to the DNA repair machinery. Indeed, comparative analysis of the proteins involved in DNA repair reveals an extreme diversity of the repair systems in bacteria and archaea/eukaryotes (35).

As an alternative to all these explanations, we hypothesize that the modern-type systems for dsDNA replication evolved independently in bacteria and in the archaeal/eukaryotic lineage. In the proposed model, the LCA did not have a replicating DNA genome and instead maintained a mixed RNA/DNA genome that had the following basic properties (Fig. 2): (i) genomic RNA was reverse-transcribed into a RNA/DNA heteroduplex by a reverse transcriptase; (ii) the RNA moiety of the RNA/DNA duplex was digested by a nuclease; (iii) the remaining ssDNA served as the template for the synthesis of a dsDNA molecule (this step can be catalyzed by the same reverse transcriptase as step 1); (iv) RNA was transcribed by a DdRp from the DNA genome (this step is the evolutionary forerunner of modern-day transcription).

This model explains the universal conservation of the core transcription machinery, the enzymes for DNA precursor biosynthesis and those components of the extant replication machinery that are orthologous and highly conserved in all forms of life, namely RNase HII and FEN1-like 5'→3' exonuclease. The role of the other universal components of the replication machinery, such as the sliding clamp, the clamp-loader, the ligase and topoisomerase I, is less obvious and they do not seem to be required for the postulated mixed genetic system to function. It is conceivable, however, that a sliding clamp and a clamp-loader functioned in the LCA to increase the processivity of reverse transcription.

This model assumes a central function for a reverse transcriptase in the replication cycle of the LCA. Moreover, the hypothetical cycle that we have inferred by comparing the cellular DNA replication machinery components strikingly resembles those of retroid viruses, particularly caulimoviruses and hepadnaviruses (55). The similarities between the retroviral replication system and that of a hypothetical ancient cellular organism have been considered by Wintersberger and Wintersberger (56). It is conceivable that present-day retroid viruses are descendants of ancient genetic elements that escaped during the reverse transcription stage of cellular replication. The existence of an astonishing variety of reverse transcribing genetic elements, both RNA- and DNA-based, in modern-day eukaryotes and bacteria is not incompatible with this idea. On the other hand, except for eukaryotic telomerases (57) and eubacterial multicopy ssDNA-related enzymes (58), reverse transcriptases are rarely encoded by cellular genomes. It appears that reverse transcriptase cannot be tolerated by DNA replication-competent cells. Once DdDps have evolved, selection would favor elimination of the reverse transcription pathway to prevent the 'backward' propagation of damage to RNA into DNA.

A notable aspect of the conservation pattern of the transcription machinery components supports this reverse transcription-based model. While the principal RNA polymerase subunits are highly conserved in the three domains of life, the subunits that are required for gene-specific transcription, such as the σ-factors in bacteria and TFIIB/TBP in archaea/eukaryotes, show no relationship beyond the generic nucleic acid binding helix–turn–helix domain (Table 1; 59). This suggests that in the LCA, the RNA polymerase might not have been used for gene-specific transcription, but rather as a 'replicative enzyme' (Fig. 2).

An important feature of the discussed model (as probably in any RNA genome model) is that the genome of the LCA consisted of multiple segments, simply because very long RNA molecules are unstable. A further attractive possibility is that circular DNA intermediates could have been formed in the LCA via mechanisms similar to those involved in the formation of circular proviruses in extant retroviruses and/or the virion dsDNA of hepadnaviruses and caulimoviruses (55). The formation and subsequent transcription of such circular dsDNA elements could have required the function of DNA ligase and topoisomerase I, respectively, thus justifying their likely presence in the LCA. Furthermore, the size of these replicons could increase via recombination, leading to an increasing demand for the sliding clamp, the clamp-loader and the topoisomerase and mounting pressure for the 'invention' of a true DNA replication system. A hint that recombination might have been actively occurring at this stage is the ubiquity and substantial conservation of RecA/RadA (the principal recombination ATPase) in all extant life forms (35). The presence of replicons of substantial size (~30 kb) at this point in evolution is suggested by the conservation in bacteria and archaea of the ribosomal protein super-operon, which encodes some of the most highly conserved proteins in all life forms, namely the ribosomal proteins and RNA polymerase subunits (60,61). In all likelihood, this super-operon has been inherited from the LCA. Thus the first, 'provirus-like' DNA molecules could have been the precursors of bacterial-size circular dsDNA replicons, probably the ancestral form for all modern-type DNA genomes. This could happen, however, only after an efficient DNA replication system came to be—according to our hypothesis, independently in bacteria and in archaea-like ancestors of modern archaea and eukaryotes.

The outlined model of a mixed (hybrid) RNA/DNA genome should be conceived of as an intermediate stage between a pure RNA genome and the current, DNA-based genetic system. Initially, autonomous (non-DNA-dependent) RdRp-mediated RNA replication might also have persisted (Fig. 2). Once RNA replication has ceased, a true hybrid genome (rather than a dual genome) has evolved in which RNA depends on DNA for its replication and DNA depends on RNA. Though cumbersome from today's (cells) point of view, in the absence of true DNA replication capabilities, this hybrid RNA/DNA genome seems to be the only way that a cell can benefit from the higher stability of DNA and its amenability to repair.

The portrait of the LCA emerging from this model has features that are similar to those proposed by other theories of early evolution, as well as unique ones. The model seems to be compatible with the notion of asynchronous 'crystallization' of different cellular systems recently discussed by Woese (62). In the postulated LCA with a mixed genetic system, the translation system is expected to be largely similar to the extant one and so are the principal aspects of transcription. Also, this organism should encode significant metabolic capabilities, including those for the synthesis of amino acids and ribo- and deoxynucleotides. In contrast, the replication system as we know it today is non-existent and the genome organization itself is not 'crystallized'. This creates potential for rapid evolution via recombination and re-assortment of genome segments.

The hypothesis of an independent evolution of DNA replication offers a parsimonious explanation for the strange assortment of apparently unrelated, homologous but not orthologous and orthologous components in the DNA replication machineries of bacteria and archaea/eukaryotes. Admittedly, this scenario cannot completely invalidate the competing hypothesis of an origin of the DNA replication machinery in the LCA followed by as yet unknown (but clearly dramatic) evolutionary events causing the observed dissimilarity. We may never know the final answer. It is conceivable, however, that sequencing of genomes from very early branchings of life, such as Korarchaeota, and determination of key protein structures that are still unresolved, such as the bacterial pol III α-subunit, the large subunits of the DdRp and the unique archaeal DNA polymerase, might shift the balance toward one or the other of these competing hypotheses.

REFERENCES

1. Baker,T.A. and Bell,S.P. (1998) *Cell*, **92**, 295–305.
2. Kornberg,A. and Baker,T. (1991) *DNA Replication*, 2nd Edn. W.H. Freeman and Co, New York, NY.
3. Mushegian,A.R. and Koonin,E.V. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
4. Edgell,D.R. and Doolittle,W.F. (1997) *Cell*, **89**, 995–998.
5. Bernad,A., Blanco,L., Lazaro,J.M., Martin,G. and Salas,M. (1989) *Cell*, **59**, 219–228.
6. Kong,X.P., Onrust,R., O'Donnell,M. and Kuriyan,J. (1992) *Cell*, **69**, 425–437.
7. Krishna,T.S., Kong,X.P., Gary,S., Burgers,P.M. and Kuriyan,J. (1994) *Cell*, **79**, 1233–1243.
8. Raghunathan,S., Ricard,C.S., Lohman,T.M. and Waksman,G. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 6652–6657.
9. Bochkarev,A., Pfuetzner,R.A., Edwards,A.M. and Frappier,L. (1997) *Nature*, **385**, 176–181.
10. Chedin,F., Seitz,E.M. and Kowalczykowski,S.C. (1998) *Trends Biochem. Sci.*, **23**, 273–277.
11. Hwang,K.Y., Baek,K., Kim,H.Y. and Cho,Y. (1998) *Nature Struct. Biol.*, **5**, 707–713.
12. Mueser,T.C., Nossal,N.G. and Hyde,C.C. (1996) *Cell*, **85**, 1101–1112.
13. Kim,Y., Eom,S.H., Wang,J., Lee,D.S., Suh,S.W. and Steitz,T.A. (1995) *Nature*, **376**, 612–616.
14. Ceska,T.A., Sayers,J.R., Stier,G. and Suck,D. (1996) *Nature*, **382**, 90–93.
15. Argos,P. (1988) *Nucleic Acids Res.*, **16**, 9909–9916.
16. Delarue,M., Poch,O., Tordo,N., Moras,D. and Argos,P. (1990) *Protein Eng.*, **3**, 461–467.
17. Ito,J. and Braithwaite,D.K. (1991) *Nucleic Acids Res.*, **19**, 4045–4057.
18. Ollis,D.L., Brick,P., Hamlin,R., Xuong,N.G. and Steitz,T.A. (1985) *Nature*, **313**, 762–766.
19. Kohlstaedt,L.A., Wang,J., Friedman,J.M., Rice,P.A. and Steitz,T.A. (1992) *Science*, **256**, 1783–1790.
20. Sousa,R., Chung,Y.J., Rose,J.P. and Wang,B.C. (1993) *Nature*, **364**, 593–599.
21. Wang,J., Sattar,A.K., Wang,C.C., Karam,J.D., Konigsberg,W.H. and Steitz,T.A. (1997) *Cell*, **89**, 1087–1099.
22. Aravind,L., Leipe,D.D. and Koonin,E.V. (1998) *Nucleic Acids Res.*, **26**, 4205–4213.
23. Brendel,V., Brocchieri,L., Sandler,S.J., Clark,A.J. and Karlin,S. (1997) *J. Mol. Evol.*, **44**, 528–541.
24. Forterre,P., Benachenhou-Lahfa,N., Confalonieri,F., Duguet,M., Elie,C. and Labedan,B. (1992) *Biosystems*, **28**, 15–32.
25. Forterre,P. (1997) *Curr. Opin. Genet. Dev.*, **7**, 764–770.
26. Tauer,A. and Benner,S.A. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 53–58.
27. Kelman,Z. and O'Donnell,M. (1995) *Nucleic Acids Res.*, **23**, 3613–3620.
28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
29. Altschul,S.F. and Koonin,E.V. (1998) *Trends Biochem. Sci.*, **23**, 444–447.
30. Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) *Nature Genet.*, **6**, 119–129.
31. Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol.*, **266**, 554–571.
32. Walker,D.R. and Koonin,E.V. (1997) *Ismb*, **5**, 333–339.
33. Tatusov,R.L., Mushegian,A.R., Boork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1996) *Curr. Biol.*, **6**, 279–291.
34. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
35. Aravind,L., Walker,D.R. and Koonin,E.V. (1999) *Nucleic Acids Res.*, **27**, 1223–1242.
36. Aravind,L. and Koonin,E.V. (1999) *J. Mol. Biol.*, **287**, 1023–1040.
37. Ishino,Y., Komori,K., Cann,I.K. and Koga,Y. (1998) *J. Bacteriol.*, **180**, 2232–2236.
38. Cann,I.K., Komori,K., Toh,H., Kanai,S. and Ishino,Y. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 14250–14255.
39. Aravind,L. and Koonin,E.V. (1998) *Nucleic Acids Res.*, **26**, 3746–3752.
40. Gorbalenya,A.E. and Koonin,E.V. (1989) *Nucleic Acids Res.*, **17**, 8413–8440.
41. Gorbalenya,A.E. and Koonin,E.V. (1993) *Curr. Opin. Struct. Biol.*, **3**, 419–429.
42. Neuwald,A.F., Aravind,L., Spouge,J.L. and Koonin,E.V. (1999) *Genome Res.*, **9**, 27–43.
43. Biswas,S.B., Chen,P.H. and Biswas,E.E. (1997) *Biochemistry*, **36**, 13270–13276.
44. Altamura,N., Groudinsky,O., Dujardin,G. and Slonimski,P.P. (1992) *J. Mol. Biol.*, **224**, 575–587.
45. Applequist,S.E., Selg,M., Raman,C. and Jack,H.M. (1997) *Nucleic Acids Res.*, **25**, 814–821.
46. DeMarini,D.J., Winey,M., Ursic,D., Webb,F. and Culbertson,M.R. (1992) *Mol. Cell. Biol.*, **12**, 2154–2164.
47. Koonin,E.V. (1993) *Nucleic Acids Res.*, **21**, 2541–2547.
48. Moser,M.J., Holley,W.R., Chatterjee,A. and Mian,I.S. (1997) *Nucleic Acids Res.*, **25**, 5110–5118.
49. Murzin,A.G. (1993) *EMBO J.*, **12**, 861–867.
50. Webster,G., Genschel,J., Curth,U., Urbanke,C., Kang,C. and Hilgenfeld,R. (1997) *FEBS Lett.*, **411**, 313–316.
51. O'Donnell,M., Onrust,R., Dean,F.B., Chen,M. and Hurwitz,J. (1993) *Nucleic Acids Res.*, **21**, 1–3.
52. Singleton,M.R., Hakansson,K., Timson,D.J. and Wigley,D.B. (1999) *Structure*, **7**, 35–42.
53. Bergerat,A., de Massy,B., Gadelle,D., Varoutas,P.C., Nicolas,A. and Forterre,P. (1997) *Nature*, **386**, 414–417.
54. Wolffe,A.P. and Hayes,J.J. (1999) *Nucleic Acids Res.*, **27**, 711–720.
55. Fields,B.N., Knipe,D.M., Chanock,R.M., Hirsch,M.S., Melnick,J.L., Monath,T.P. and Roizman,B. (1990) *Virology*, 2nd Edn. Raven Press, New York, NY.
56. Wintersberger,U. and Wintersberger,E. (1987) *Trends Genet.*, **3**, 198–202.
57. Cech,T.R., Nakamura,T.M. and Lingner,J. (1997) *Biokhimiia*, **62**, 1202–1205.
58. Inouye,S. and Inouye,M. (1995) *Virus Genes*, **11**, 81–94.
59. Makarova,K.S., Aravind,L., Galperin,M.Y., Grishin,N.V., Tatusov,R.L., Wolf,Y.I. and Koonin,E.V. (1999) *Genome Res.*, **9**, 608–628.
60. Koonin,E.V. and Galperin,M.Y. (1997) *Curr. Opin. Genet. Dev.*, **7**, 757–763.
61. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) *Trends Biochem. Sci.*, **23**, 324–328.
62. Woese,C. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 6854–6859.
63. Murzin,A.G. (1995) *Nature Struct. Biol.*, **2**, 25–26.
64. Ban,C. and Yang,W. (1998) *Cell*, **95**, 541–552.
65. Doherty,A.J., Serpell,L.C. and Ponting,C.P. (1996) *Nucleic Acids Res.*, **24**, 2488–2497.
66. Bork,P., Hofmann,K., Bucher,P., Neuwald,A.F., Altschul,S.F. and Koonin,E.V. (1997) *FASEB J.*, **11**, 68–76.
67. Shen,B., Qiu,J., Hosfield,D. and Tainer,J.A. (1998) *Trends Biochem. Sci.*, **23**, 171–173.
68. Barnes,M.H., Leo,C.J. and Brown,N.C. (1998) *Biochemistry*, **37**, 15254–15260.