J Mol Evol (2002) 54:763–773
DOI: 10.1007/s00239-001-0078-x

JOURNAL OF MOLECULAR EVOLUTION

Evolution of DNA Polymerase Families: Evidences for Multiple Gene Exchange Between Cellular and Viral Proteins

Jonathan Filée,[1] Patrick Forterre,[1] Tang Sen-Lin,[2] Jacqueline Laurent[1]

[1] Institut de Génétique et Microbiologie, Bât. 409, CNRS UMR 8621, Université Paris-Sud XI, F-91405 Orsay Cedex, France
[2] Department of Microbiology and Immunology, The University of Melbourne, Parkville, 3052, Victoria, Australia

Received: 25 March 2001 / Accepted: 27 November 2001

Abstract. A phylogenetic analysis of the five major families of DNA polymerase is presented. Viral and plasmid sequences are included in this compilation along with cellular enzymes. The classification by Ito and Braithwaite (Ito and Braithwaite 1991) of the A, B, C, D, and X families has been extended to accommodate the "Y family" of DNA polymerases that are related to the eukaryotic RAD30 and the bacterial UmuC gene products. After analysis, our data suggest that no DNA polymerase family was universally conserved among the three biological domains and no simple evolutionary scenario could explain that observation. Furthermore, viruses and plasmids carry a remarkably diverse set of DNA polymerase genes, suggesting that lateral gene transfer is frequent and includes non-orthologous gene displacements between cells and viruses. The relationships between viral and host genes appear very complex. We propose that the gamma DNA polymerase of the mitochondrion replication apparatus is of phage origin and that this gene replaced the one in the bacterial ancestor. Often there was no obvious relation between the viral and the host DNA polymerase, but an interesting exception concerned the family B enzymes: in which ancient gene exchange can be detected between the viruses and their hosts. Additional evidence for horizontal gene transfers between cells and viruses comes from an analysis of the small damage-inducible DNA polymerases. Taken together, these findings suggest a complex evolutionary history of the DNA replication apparatus that involved significant exchanges between viruses, plasmids, and their hosts.

Key words: DNA polymerase — Non-orthologous gene displacement — Horizontal gene transfer — Phylogeny — Virus — Phage — Plasmid

Introduction

During the past decade a large number of DNA polymerase sequences from all three domains of life (Archaea, Bacteria, and Eukarya) have been entered in the databases. These gene sequences can be arranged into families or subfamilies and various phylogenies have been proposed for these enzymes (Braithwaite and Ito 1993; Edgell et al. 1998; Huang and Ito 1999). Ito and Braithwaite, in a now classic paper, classified all the DNA polymerases sequences into four families (A, B, C, and X) (Ito and Braithwaite 1991). This DNA polymerase phylogeny is dated and in the light of the enormous amount of sequence data accumulated in the last ten years requires a re-examination. In addition, DNA polymerases are the ideal phylogenetic markers to study ancient relationships between cellular and viral genes, since they offer the greatest number of viral and cellular homologs.

There is considerable diversity among DNA polymerases and their functions are not all identical. Recently, a new family of DNA polymerases has been discovered in the Euryarchaea (Ishino et al. 1998) and many small

Correspondence to: J. Laurent; *email:* Jacqueline.Laurent@igmors.u-psud.fr

Table 1. The world of DNA polymerase: families, distribution, activities, and particularities

Polymerase family	Family A	Family B	Family C	Family D	Family X	Family Y Rad30/DIN/UmuD
Repartition	-Bacteria	-γ-Proteobacteria	-Bacteria		-Bacteria	-Bacteria (DinX, DinB, UmuD)
	-Mitochondrion (γ)					
	-Metazoa, Plants (mus308)	-Eukaryota (α, δ, ζ, ϵ)			-Eukaryota (μ, β, TdT)	-Eukaryota (Rad30, ι, DinB, REV1)
		-Archaea		-Euryarchaea	-Archaea	
		-Plasmids, Viruses,	-Cryptic phages		-Viruses	-Archaea (Dbh)
	-Phages	Phages	-Plasmids			-Plasmids
Associated Activity	3′–5′ Exonuclease	3′–5′ Exonuclease	3′–5′ Exonuclease	3′–5′ Exonuclease	5′ phosphatase (β)	
	5′–3′ Exonuclease	Primase (α)				
Particularity		2 categories:			Template-independent polymerase	
		-Protein-primed				
		-RNA-primed				
		Eukaryotic polymerase ζ performs translesional reparation				

DNA polymerases have been described in Bacteria and Eukaryotes (Goodman and Tippin 2000) that are involved in mutagenic repair. The characteristics of the six major groups of DNA polymerases are summarized in Table 1. DNA polymerases belonging to families B, C, and D appear to be involved in chromosomal replication, while A-type DNA polymerases replicate the mitochondrial DNA. Only A and B type DNA polymerases replicate viral genomes, but the repair of DNA seems to involve DNA polymerases from all the families (Hübscher et al. 2000). On the basis of the comparison of their amino acid sequences, the DNA polymerase families all seem to be unrelated. However, the A and B families share some common biochemical and structural features (Steitz 1999), particularly in the domain encoding their 3′–5′ exonuclease (Zhu and Ito 1994). The strong conservation of amino-acid motifs in the three "exo boxes" make it likely that at least the exonuclease domains of the A, B, and C DNA polymerases families are homologs. It is unclear, however, if the common features observed in their polymerization domains share a homologous origin or not. In some of the polymerases of the ABC families, a separate polypeptide encodes the exonuclease domain (e.g. bacterial DNA polymerase III) whereas other polymerases lack this exonuclease activity (e.g. the Taq polymerase). Although some polymerases form large protein complexes (especially those involved in DNA replication) others are monomers or dimers of proteins with only one or a few domains.

Genetic and biochemical studies have established that a DNA polymerase of the C family is the principal DNA replication enzyme in bacteria (Kornberg and Baker 1992). Surprisingly, a gene encoding for a C-type DNA polymerase is absent from both the Archaea and Eukarya (Table 1). In Eukarya, several polymerases belonging to the B family (α, δ, ϵ) have been implicated in chromosomal DNA replication (Hübscher et al. 2000). The identity of the archaeal replicase is unknown, but from the distribution of DNA polymerase genes in the genomes of these organisms, it seems likely that the Crenarchaea use a B DNA polymerase for replication process and the Euryarchaea use either a B or a D polymerase, or both (Cann and Ishino 1999). This distribution illustrates the striking differences in the replication machinery of the three biological domains, although the archaea and Eukaryotes have some similarities. Several hypotheses have been proposed to explain this observation. Koonin and his co-workers have suggested that the LUCA (the Last Universal Common Ancestor) had an RNA genome and that DNA replication independently evolved twice, once in the bacterial lineage and again in the archaeal/eukaryotic lineage (Leipe et al. 1999). Another hypothesis postulates that the DNA replication machinery of LUCA was rather flexible in its constituents and that none of them was universally conserved, or alternatively that these sequences have diverged so much that the original homology is no longer detectable (Edgell and Doolittle 1997). Finally, a third hypothesis proposed that non-orthologous gene displacement was the explanation for the sequence diversity (Forterre 1999). In this case, unrelated or paralogous proteins are responsible for the same critical functions in different species (Koonin et al. 1996). Forterre suggested that plasmids and viruses might often be the donors of the non-orthologous replication genes that replaced the ancestral cellular versions (Forterre 1999). The study of DNA polymerases evolution seems particularly appropriate to test the validity of the latter hypothesis, since all DNA polymerase families

possess plasmid or viral members. Also because DNA polymerases are involved in numerous different functions (Table 1). A good example of non-orthologous displacement of cellular gene by a viral version is the case of RNA polymerase of mitochondrion which is apparently derived from a bacteriophage T3/T7-type RNA polymerase rather than a bacterial enzyme (Gray and Lang 1998). Villarreal (Villarreal 1999) has discussed a hypothesis that the DNA polymerase of eukaryotes had its origins in eukaryotic DNA viruses.

We have performed a comprehensive phylogenetic analysis of DNA polymerases with the objective of clarifying the relationships and critically testing the hypothesis of non-orthologous gene displacement by genes of viral or plasmid origin. To do such a global analysis, we have extensively screened public databases using multiple query sequences to retrieve as many viral, plasmid and cellular DNA polymerases sequences as possible. We have also included in this study several unpublished DNA polymerases sequences obtained from recently characterized archaeal viruses HF1 and HF2 (Nuttall and Dyall-Smith 1995) and from T4-related phages (RB49 and RB69). Thus, this analysis is the first to include viral and cellular enzymes from each of the three biological domains. Our study reveals that mitochondrial DNA polymerases of the A family are only distantly related to their bacterial homologs and may have originated from a bacteriophage (as it is the case for the RNA polymerase) rather than from the α-proteobacterial ancestor of mitochondria. Additional evidence was found for horizontal transfer between viruses and plasmids and their hosts. Most of these examples concerned the small, damage-inducible polymerases of the mutagenic repair pathway and appeared to be rather recent.

The B-type DNA polymerases of the archaeal virus HF2 and HF1 appears very interesting because they are closely related to the DNA polymerase of their host but are only distantly related to the other Archaeal DNA polymerase. This suggests a rather recent horizontal gene transfer of the halovirus-type DNA polymerase in the genome of their host. Finally, we also obtained indications of a very ancient gene transfer between eukaryotic cells and their viruses but without being able to define the direction of this transfer.

This phylogenetic study of DNA polymerases clarifies the relationships within each polymerase family and revealed that the relationships between the viral and host genes are very complex, probably reflecting a very long evolutionary history dating back to LUCA or, perhaps, even earlier.

Materials and Methods

BLAST (Altschul et al. 1990) or PSI BLAST (Altschul and Koonin 1998) search with a threshold probability value for inclusion in the first step of iteration of 0.002, identified all sequences belonging to each family of DNA polymerase available in data banks. The program ALIBABA (Philippe Lopez, personal communication) allowed us to retrieve all sequences automatically and to write them into a MUST-compatible file (Philippe 1993). Samples were completed with preliminary sequence data obtained from the Institute for Genomic Research website at http://www.tigr.org.

Because of the high degree of divergence of the regions surrounding the conserved domains, the alignment of these sequences was carried out in two steps. Initially, sequences thought to be orthologous were aligned with each other (for example all the mitochondrion A-family or all the eukaryotic α polymerase) using CLUSTAL W (Thompson et al. 1994) and refined manually with the help of the ED program of the MUST package version 3.0 (Philippe 1993). Then, separated alignments of orthologs were combined and aligned by hand with the help of the BLAST output. Concerning the families A, B, and C of DNA polymerase, we used as reference the alignment proposed by Braithwaite and Ito (Braithwaite and Ito 1993). All the alignments can be found at http://www-archbac.u-psud.fr/Projects/dnapol/Ali_debut.htm. Positions that could not be unambiguously aligned were excluded from the analysis, and gaps were removed. Phylogenetic trees were constructed with the maximum-likelihood (ML), maximum parsimony (MP), and distance-based methods, running the programs PROTML (Adachi and Hasegawa 1996) version 2.3, PAUP (Swofford 1993) version 3.1, and NJ in the MUST package (Philippe 1993) version 3.0, respectively. MP trees were obtained by a heuristic search. NJ trees were obtained without any distance correction (p-distance) and bootstrap proportions (BP) were calculated by analysis of 1000 replicates using the NJBOOT program of the MUST package (Philippe 1993) version 3.0.

Due to the high number of taxa used, the exhaustive search of the ML trees were performed in constraining some nodes. We used the NJ and PAUP trees topology to contribute to constrain taxa. The model of amino acid substitution used was JTT-F (Jones et al. 1992). Bootstrap values were calculated using the RELL method with the BOOTML program (Philippe, personal communication) on the 1000 best trees.

Results

Phylogeny of Family A Polymerases

The prototype of DNA polymerases family A is the *E. coli* DNA polymerase I (Kornberg's polymerase). A PSI-BLAST-search performed with *E. coli* Pol I sequence as the query, retrieved the bacterial homologs, followed by several bacteriophage DNA polymerases of both Gram positive or Gram negative hosts and the N-terminal domain of a novel eukaryotic nuclear DNA polymerase/helicase (Harris et al. 1996). The mitochondrial gamma-type DNA polymerases, which are also encoded by nuclear genes, were retrieved using a PSI-BLAST after one step of iteration. A BLAST search seeded with the sequence of *Homo sapiens* Mus308 gene product (also called DNA polymerase eta by Burtis and Harris in 1997) retrieved first the sequence of a proteobacterium (*Rhodothermus sp.*), suggesting that the eukaryotic Mus308 gene product could be of mitochondrial origin.

The alignment revealed that the exonuclease domains were very poorly conserved outside of the three exo boxes. So we used only the polymerase domains to construct the phylogenetic trees. We finally retained for the analysis 122 positions for 65 taxa. We first performed a

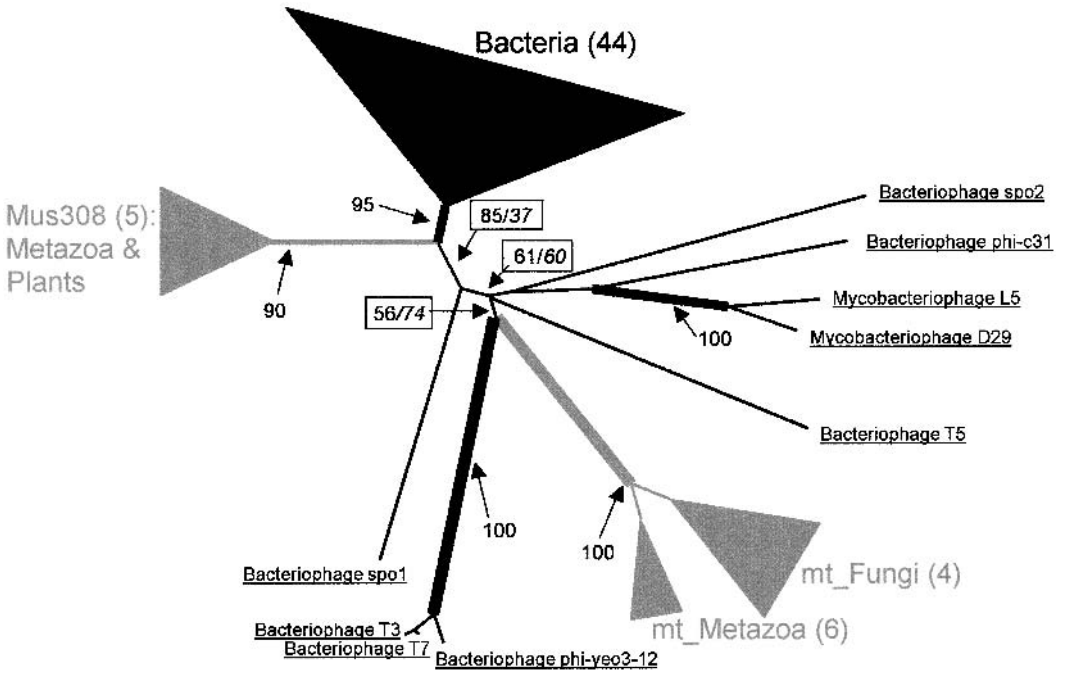

Fig. 1. ML phylogeny of the family A of DNA polymerases. Thick branches indicate constrained groups during the analysis (see text). Viruses names are underlined. Figures in italics indicate ML-bootstrap values and roman ones NJ-bootstrap values. In frames are indicated the bootstrap values obtained by both methods for the same branch. For constrained clusters, there is only the NJ-bootstrap value. Branch length is proportional to distance. Numbers in brackets indicated the number of taxa.

NJ phylogeny. The result of that analysis helped us to constrain some nodes for the ML analysis, giving the unrooted tree presented in Fig. 1. To accelerate the ML analysis, only eight sequences of proteobacteria were used. The results of these phylogenetic analyses confirmed the specific relationships between DNA polymerases from bacteria and eukaryotic nuclear DNA polymerase/hélicase first suggested by BLAST. Indeed, they were closely related in both ML and NJ trees, showing a grouping of two monophyletic groups, each containing either the eukaryotic sequences or the bacterial ones. Phylogenetic analysis also confirmed that mitochondrial DNA polymerase gamma and bacteriophage DNA polymerases are only distantly related to their bacterial homologs. Bacteriophage DNA polymerases all appear very remote from each other (except for the group of T3 related bacteriophage and two Mycobacteriophages) with no specific relationships between bacteriophages infecting the same host (see Spo1/Spo2 or T3/T5). Interestingly, mitochondrial DNA polymerase gamma turned out to be sister group to bacteriophages T3/T7/Phi-yeo3-12 DNA polymerases. This grouping was found whatever the reconstruction method used with variable bootstrap values (56% in NJ and 74% in ML). Although this grouping may result from long branch attraction, this is reminiscent of the mitochondrial RNA polymerase of the T3/F7 type (Gray and Lang 1998), suggesting that several informational proteins in mitochondria originated from the same phage of the T3/F7 family by non-orthologous displacement. Finally, a plausible hypothesis to explain the likely monophyly of eukaryotic Mus308 gene product and bacterial sequences is that the gene encoding the ancestral bacterial A-type DNA polymerase migrated to the nucleus where it fused with a helicase module to give rise to Mus308 gene product. However, as only unrooted trees can be displayed, the phylogenetic relationships revealed in this study must be interpreted with caution.

Phylogeny of Family B DNA Polymerases

Although DNA polymerases of the B family are rare in Bacteria (only in Gamma proteobacteria) the prototype of this family is *E. coli* DNA polymerase II (Ito and Braithwaite 1991). Two categories of DNA polymerases can be distinguished according to their mechanistic properties: those which use either DNA or RNA as primers as in the case of most other DNA polymerases (RNA/DNA-priming type) and those which use a unique protein-priming mechanism (protein-priming type). The RNA-priming class is present in a wide range of organisms including Archaea, Eukaryota, Enterobacteria and many eukaryotic and prokaryotic viruses, whereas protein-priming subfamily comprises the sequences of DNA polymerases encoded by eukaryotic linear plasmids of mitochondrion (principally some Fungi and two Plants), of many bacteriophages, and of eukaryotic Adenovirus.

Interestingly, using *E. coli* DNA polymerase II as query sequence, we were able to retrieve all sequences

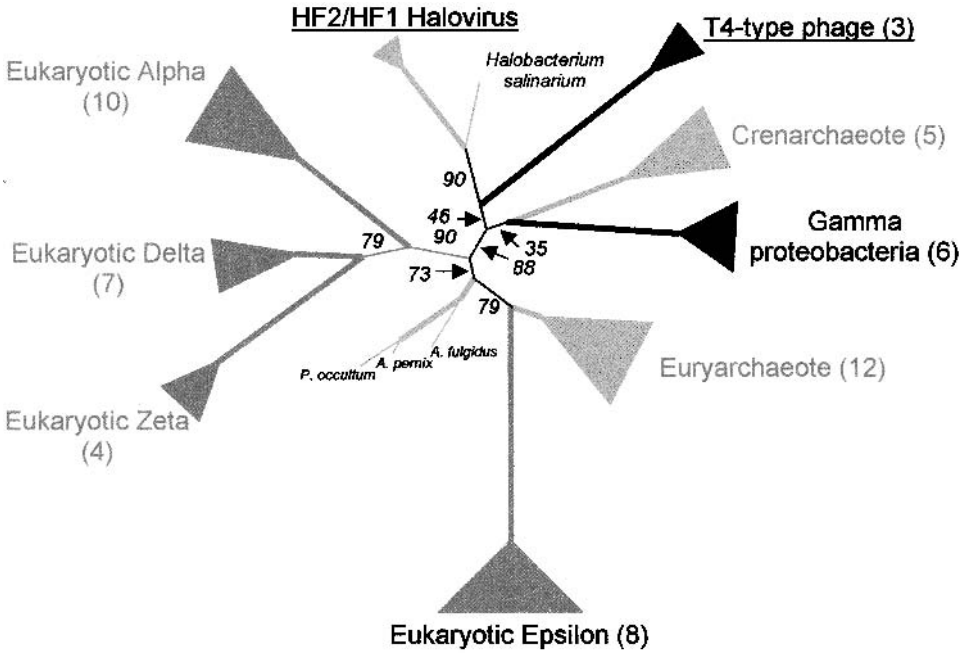

Fig. 2. ML phylogeny of the family B of DNA polymerases (DNA/RNA primed). Thick branches indicate constrained groups during the analysis (see text). Figures in italics: ML-bootstrap values. Viruses names are underlined. Numbers in brackets indicate the number of taxa. Branch length is proportional to distance. Only a restricted number of taxa are used (see text). The complete phylogeny of the family is available at http://www-archbac.u-psud.fr/Projects/dnapol/Ali_debut.htm.

known to belong to the RNA/DNA-priming class but no sequence known to correspond to protein-priming type. A second iteration of the PSI-BLAST search was necessary to retrieve, with high *E* value ($10^{-10} < E < 10^{-3}$), the protein-priming DNA polymerases. Similarly, using protein-priming DNA polymerase sequences as query, we never retrieved RNA/DNA priming polymerases. This suggests to divide the family B DNA polymerase in two subfamilies, the RNA/DNA-priming subfamily and the protein-priming subfamily. This also suggested that a simple BLAST analysis should be able to predict if a new polymerase of the B family is of the RNA/DNA or of the protein priming type.

When PSI-BLAST searches were performed with sequences of DNA polymerases from the two haloarchaeophages HF1 and HF2, it appeared that these enzymes belonged to the DNA/RNA priming type with highly significant *E* value ($<10^{-105}$). We retrieved first the DNA polymerase B1 from *Halobacterium sp.* followed by some archaeal sequences (*Archaeoglobus fulgidus, Pyrodictium occultum*) and by the sequences of the RB69/T4 phage family (*E* value close to 10^{-25}).

The primary sequences of the protein-priming subfamily of DNA polymerase B presented very few similarities and we were unable to identify a sufficient number of conserved residues to make a convincing alignment. In contrast, we were able to perform a phylogenetic analysis of the RNA/DNA priming subfamily using the most conserved regions: the exonuclease domains I and II and polymerase domains I to VII (Edgell et al. 1998). The alignment of these regions yielded to 128 usable positions in 91 taxa. Unrooted MP and NJ global analysis yielded poorly resolved phylogenies that were not congruent between the two methods. Only one large group composed of the eukaryotic polymerase δ and the sequences of Herpes virus, Phycodnavirus (bootstrap value of 80%) as well as Ascovirus and Iridovirus sequences with lower statistical support, were clearly distinguished (the NJ phylogeny of the family is available at http://www-archbac.u-psud.fr/Projects/dnapol/NJ-polB.htm). Numerous short signatures also supported this clustering; the larger one can be seen at http://www-archbac.u-psud.fr/Projects/dnapol/Signature.htm with a representative set of sequences. This grouping is in agreement with the results of Villareal and DeFilippis (Villarreal and DeFilippis 2000).

To investigate further the relationships among the DNA/RNA priming class of the B family, we performed an ML analysis using a limited number of taxa. We eliminated the sequences of the eukaryotic viruses and some highly divergent paralogs of Archaea (genus *Sulfolobus,* and species *Archaeoglobus fulgidus,* and *Halobacterium sp.*). To reduce the calculation time in ML, the monophyly of several groups evidenced by NJ were constrained. The resulting ML phylogeny (Fig. 2) can be divided in three large groups. A first group contained the eukaryotic α, δ, and ζ (rev3) paralogs with high bootstrap values (90% in ML). A second group was composed of the eukaryotic polymerase ϵ and diverse archaeal genes including all the Euryarchaeote genes

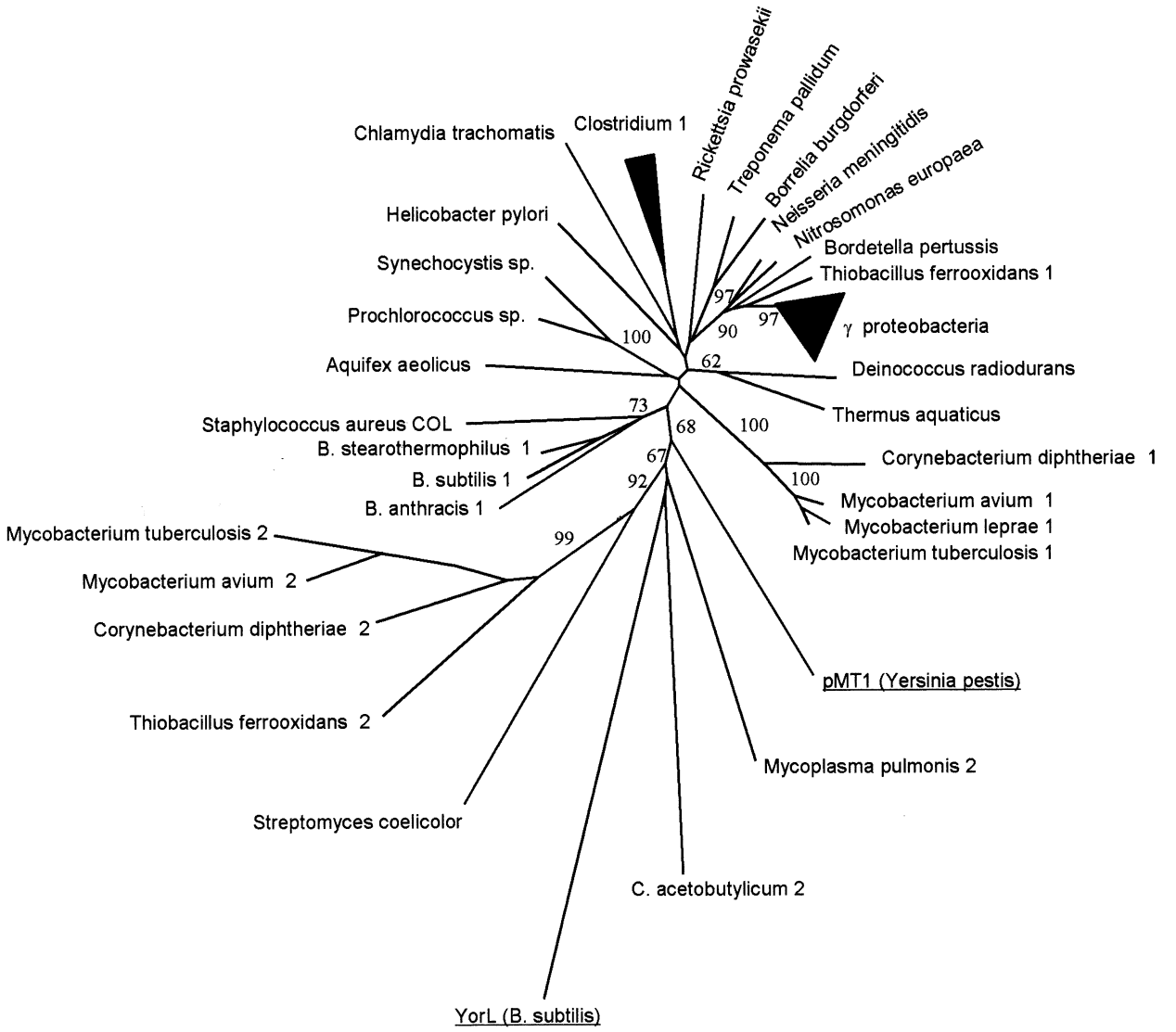

Fig. 3. NJ phylogeny of the family C of DNA polymerases. Figures indicate NJ-bootstrap values. Names of viruses are underlined. The hosts of the viruses and plasmids are indicated in brackets. Branch length is proportional to distance.

(excepted those of *Halobacterium sp.*) and three Crenarchaeote genes of *Aeropyrum fulgidus, A. pernix,* and *Pyrodictium occultum.* This node is moderately supported in ML (73%). Finally, a third group consisted of all the gamma proteobacteria B-type DNA polymerases, the sequences of the remaining Crenarchaeote polymerase, as well as those of the T4-related phages, the HF1-HF2 halovirus and the sequence of the Euryarchaeote *Halobacterium sp.* The bootstrap support for this cluster is high in ML (88%). It is remarkable that the sequences of the halovirus HF1 and HF2 are closely related to the sequence of their host *Halobacterium sp.* with high bootstrap support (90% in ML and 94% in NJ) but this group is only distantly related to the other euryarchaeal DNA polymerases. The trees also show that the HF2-HF1/*Halobacterium sp.* cluster is sister group to the T4 phage family but with low bootstrap support in ML (46%), higher in NJ (71%).

Phylogeny of C-type DNA Polymerases

The prototype of the C family is the *E. coli* DNA polymerase III. A BLAST search with the *E. coli* Pol III sequence as the query retrieved a large number of bacterial DNA polymerases but no eukaryotic or archaeal sequence. We also got a plasmid sequence from *Yersinia pestis* as well as a sequence from a cryptic prophage in *Bacillus subtilis.* Aligning the recovered sequences indicated that 176 residues could be retained in 47 taxa. The unrooted phylogenetic tree constructed from this data (Fig. 3) was roughly similar to that obtained by Huang and Ito (1999). The sequences of the plasmid and the prophage branched off close to a group of divergent sequences from the Gram-positive bacteria from the species *Thiobacillus ferrooxidans.* These genes could be either the consequence of gene duplications or the results of lateral gene transfers possibly involving viruses. The

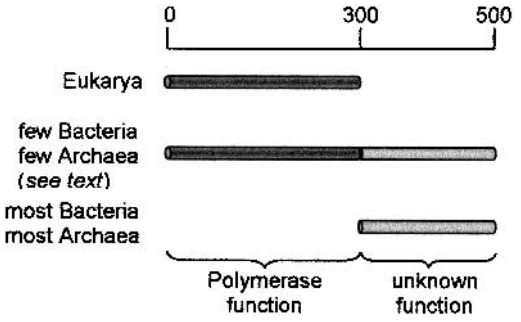

Fig. 4. Schematic representation of the modular structure of the family X DNA polymerases. The segments used in the phylogenetic analysis are indicated in green. The numbers refer to the amino acid positions.

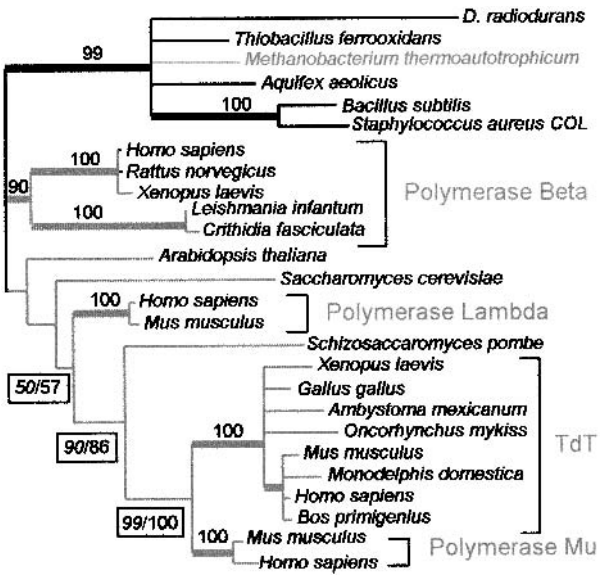

Fig. 5. ML-phylogeny of the family X of DNA polymerases. Thick branches indicate constrained groups during the analysis (see text). Figures in italics indicate ML-bootstrap values and roman ones NJ-bootstrap values. In frames are indicated the bootstrap values obtained by both methods for the same branch. Branch length is proportional to distance.

fast evolution of these sequences in this group probably results in a long branch attraction artefact (Felsenstein 1978; Philippe and Laurent 1998).

Phylogeny of X-type DNA Polymerases

The family X of DNA polymerases belongs to the large superfamily of nucleotidyltransferase which includes a large variety of molecules (Aravind and Koonin 1999). A PSI-BLAST search with the sequence of the human terminal deoxynucleotidyl transferase (TdT) retrieved all the sequences available for TdT. This molecule seemed to be present only in the vertebrates. We also found sequences of DNA polymerase μ (Eukaryote) with highly significant *E* values (close to 10^{-95}). With lower *E* value we retrieved the sequences of eukaryotic DNA polymerases β and λ ($10^{-23} < E < 10^{-15}$). After one iteration, we retrieved the sequence of the bacterium *Aquifex aeolicus,* and the Archaeon *Methanobacterium thermoautotrophicum* with an intermediate *E* value (10^{-15}) in which the motifs forming the active site for nucleotidyl transfer are present (Smith et al. 1997). The alignment of these sequences showed that only the first three hundred amino acids of the two prokaryotic sequences were homologous with the eukaryotic ones (Fig. 4). Moreover, BLAST searches seeded with the *Aquifex aeolicus* sequence showed that only a small set of species possessed this N-terminal part of the molecule (genus *Staphylococcus, Bacillus, Thiobacillus, Deinococcus,* and *Methanobacterium*). It seems that the N-terminal domain is absent in most bacterial and archaeal sequences but the C-terminal part of the prokaryotic molecule is widely represented in Archaea and Bacteria although it is lacking in Eukaryotes (a BLAST search with the last two hundred amino acids of the *Aquifex aeolicus* sequence failed to identify any eukaryotic homologue).

All the eukaryotic sequences and the N-terminal part of the bacterial sequences were aligned. Twenty-seven taxa were used to construct an exhaustive ML phylogenetic tree based on the 178 most conserved positions of molecules (Fig. 5). Clearly bacterial sequences were monophyletic and these were used to root this tree. Two types of eukaryotic paralogs seem to be present in a wide variety of Metazoa (DNA polymerase Beta and TdT) whereas the DNA polymerases Mu and Lambda are only present in Mammals. But we cannot rule out a sequence sampling bias in this distribution because only few eukaryotic genomes are presently fully sequenced. Phylogenetic analysis with high bootstrap support (99% in ML and 100% in NJ) indicates a close relationship between TdT and DNA polymerase Mu. The phylogenetic positions of the fungi sequences of *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* and of the plant *Arabidopsis thaliana* are less clear. Fungi are paraphyletic, the sequence of *S. pombe* probably belongs to the TdT/Mu group but the position of *S. cerevisiae* is ambiguous.

Phylogeny of Y-type DNA Polymerases (UmuC/DinB Super-Family)

Recently, several new DNA polymerases and a deoxycytidyl-transferase (Rev1) that are involved in bypassing the stalling of replication forks at DNA lesions have been discovered in organisms that range from *E. coli* to human. Most of them are error-prone, such as UmuC, DinB and Rev1, Pol is (L) whereas Rad30 is error free (Johnson et al. 1999). To identify members of this new family the sequence UmuC of *E. coli* was used to search (BLAST) the sequence databases. This retrieved a large number of damage-inducible proteins (UmuC, DinX, DinB) from Prokaryotes and Eukaryotes. We also re-

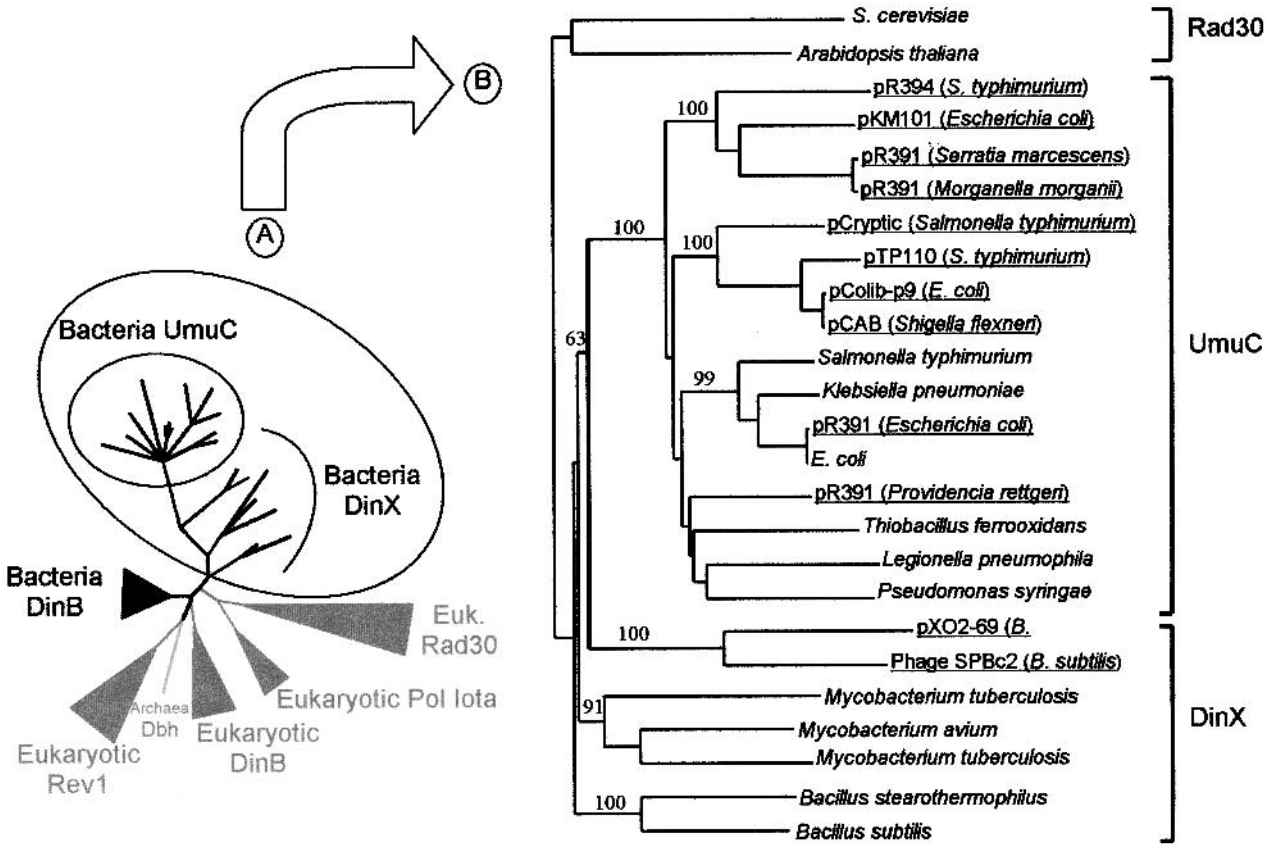

Fig. 6. Phylogeny of the family Y of DNA polymerases. **A** ML-phylogeny of the family Y of DNA polymerase. **B** NJ-phylogeny of the paralogous UmuC and DinX of the family Y of DNA polymerase. Plasmid and virus names are underlined. The hosts of the viruses and plasmids are indicated in brackets. Figures indicate NJ-bootstrap values. Branch length is proportional to distance.

trieved the sequences of the eukaryotic gene products RAD30, RAD30B, and REV1. Interestingly, we found a single archaeal sequence (Dbh gene from *Sulfolobus solfataricus*), some plasmid sequences of gamma proteobacteria, one plasmid sequence of *Bacillus subtilis* and finally a sequence of a *Bacillus* phage. All the sequences were retrieved with *E* values better than 10^{-10}. These DNA polymerases and the deoxycytidyl-transferase Rev1 can thus be considered to be members of a single new DNA polymerase family that we propose to name family Y.

For the phylogenetic analysis of this family we retained the 134 most conserved positions. The resulting ML tree is presented in Fig. 6A. Figure 6A shows the restricted NJ-phylogeny of the paralogous UmuC and DinX rooted with the eukaryotic Rad30 genes. Three major groups were detected. Eukaryotic polymerase ι and RAD30 gene product group with an especially high bootstrap proportion in ML (100%). We also observed a cluster including eubacterial and eukaryotic DinB, Rev1, and *Sulfolobus solfataricus* Dbh but with low bootstrap support (<50%). Finally, all the bacterial sequences belonging to the UmuC class and DinX class of molecules form a monophyletic group with diverse sequences of plasmids and bacteriophages. It seems that only γ-proteobacteria, Gram positive bacteria (both low and high GC groups), and *Thiobacillus ferrooxidans* possess this kind of gene, indicating multiple events of gene loss or acquisition in the other groups of bacteria. Within this group, it appeared clearly that phage and plasmid sequences were intermixed with bacterial sequences, indicating multiple independent horizontal gene transfers between viruses, plasmids, and their host.

Discussion

We have performed a phylogenetic analysis of all the families of DNA polymerases except the D-type that are present only in Euryarchaeota. Our objective was to obtain a better understanding of the evolutionary relationships within each family. Our data indicate that the general classification by Ito and Braithwaite (1991) is still valid, despite the dramatic increase in the number of sequences available. We have extended it to accommodate the "Y family" of DNA polymerases that are related to the eukaryotic RAD30 and the bacterial UmuC gene products, and we propose to split the B family into two subfamilies, the RNA/DNA primed and the protein primed B subfamilies.

Families B, X, and Y have members in all three of the

biological domains. We previously remarked that the presence of B family DNA polymerase in all three domains suggested the presence of an ancestral DNA polymerase in LUCA (Forterre 1992). On closer re-examination however, it turns out that the B family enzymes are restricted, among Bacteria, to proteobacteria. Similarly, a single archaeal DNA polymerase is present in the X and Y families. For family A polymerases, there is evidence for lateral gene transfer between viruses, bacteria and eventually to eukaryotes by the endosymbiosis of the mitochondrion (see below). If we had assumed that the original genes were present in the LUCA, one might have supposed a generalized early gene loss in the archaeal domain (families X and Y) or the bacterial domain (family B) followed by a lateral gene transfer. It is more parsimonious to suggest that these DNA polymerases were not present in the LUCA and have been introduced subsequently by gene transfer in a domain in which they were not originally present. Thus, the parsimonious explanations for the present distribution of the different DNA polymerase families between the three domains are either that the LUCA had no DNA polymerase, in agreement with the suggestion by Leipe et al. (1999), or that the primitive, ancestral polymerase sequence was lost in one or two domains and replaced by polymerases of other families (Forterre 1999).

DNA polymerases of family A appear to be strictly restricted to bacteria and eukarya. Proteins have been annotated as DNA polymerase of family A in the complete genome sequences of the archaea *Archaeoglobus fulgidus* and *Halobacterium* NRC (Ng et al. 2000). However, close inspection reveals that these proteins, which are present in other archaeal and bacterial genomes, are probably not DNA polymerases but correspond to a protein of unknown function that has been fused to the N-terminus of bacteriophage SPO1 DNA polymerase (A family) (data not shown). According to our analysis, DNA polymerases of the family A probably originated either in Bacteria or bacteriophages and were later transferred between viruses, bacteria and eventually to eukaryotes by the endosymbiosis of the mitochondrion (see below). In contrast, phylogenetic analyses did not suggest a clear hypothesis for the unusual distribution of DNA polymerases X and Y in bacteria and eukarya. Finally, some families are present only in one domain: as family C in the bacteria and family D in the archaea. These could have appeared in their domains subsequent to their divergence. As previously suggested (Forterre 1999; Villarreal and DeFilippis 2000), viral DNA polymerases could have been the source of new cellular polymerises. Indeed, viral and plasmid genes encoding for DNA polymerases are widespread in several polymerase families (A, B, Y, and to a lesser extend C). Our phylogenetic analyses argue for multiple gene exchanges occurring between viruses, plasmids and hosts in almost all these families. Some are recent (especially in family Y), but other are clearly ancient and in those cases, phylogenetic analysis cannot easily distinguish between transfers from viruses to cells or from cells to viruses (see below). The promiscuous transfer of the viral or plasmid sequences to cellular organisms could thus possibly explain the puzzling distribution of each family of DNA polymerases in the three domains of life.

Nonorthologous Displacements and Gene Fusion in the A Family

The phylogeny of DNA polymerases of the A family appears especially interesting in this context. It was known for a long time that the enzyme replicating mitochondrial DNA, DNA polymerase gamma, was a member of the A family. This suggested that it originated from the DNA polymerase I (A family) of the alphaproteobacterium which gave rise to the mitochondrion and not from the ancestral replicase (family C) of this organism. However, our data show that mitochondrial DNA polymerase γ is only distantly related to bacterial DNA polymerase I and that it branches instead with T3/T7 phage DNA polymerases. This grouping might result from a long branch attraction artefact (LBA) since the branches of phages and of DNA polymerase γ are much longer than those of the bacterial DNA polymerase and of the eukaryotic Mus308 gene products. However, since the mitochondrial RNA polymerase clearly belongs to the same family as T3/T7 RNA polymerases, it is tempting to suggest that both the transcription and replication enzymes of mitochondria have been affected by nonorthologous displacement involving proteins originating from the same T3/T7 related virus.

Moreover, we failed to identify any mitochondrial DNA polymerase belonging to family A in Plants. The unique DNA polymerase that we found in the mitochondria of *Betta vulgaris, Zea mais,* and of the red algae *Porphyra purpurea* are of the plasmid protein-primed B-type. This suggests, either that the ancestral mitochondrial replicase of the C family was directly replaced by a plasmid B-type DNA polymerase in Plants, or that, in a second displacement step, the T3/T7 phage-type mitochondrial DNA polymerase was replaced by the B-type polymerase. We favor the two-step process because the RNA polymerase of plants is of phage-type, suggesting that the replacement of the ancestral replication and transcription machineries of mitochondria by phase type enzymes occurred only once before the split between Plants and other eukaryotes.

Since BLAST search data and phylogenetic analyses of the polymerase domains of family A indicated that eukaryotic Mus308 gene products and bacterial DNA polymerases A are closely related, the eukaryotic Mus308 gene product could be the descendant of the ancestral mitochondrial DNA polymerase I. In conflict with this hypothesis, the Mus308 sequences is not a sister

group of proteobacteria in our phylogenetic analysis but branches from the base of the bacteria's cluster. However, this could be due to an increased mutation rate, resulting in the fusion of an helicase module followed by a functional shift.

Ancient Gene Transfer and Nonorthologous Displacement in the B Family

The B family of DNA polymerases contains an unusual number of viral and plasmid representatives. Phylogenetic arguments and many molecular signatures suggest that the B-type DNA polymerase sequences of Herpesvirus, Phycodnavirus, Ascovirus, and Iridovirus form a clade with the sequences of eukaryotic DNA polymerase δ. Using a similar analysis, Villarreal and De Filippis (2000) proposed that the Eukaryotic DNA polymerase δ originated from a virus. Indeed, eukaryotic DNA polymerase δ also emerges from a cluster of viral sequences in our analysis (not shown, see http://www-archbac.u-psud.fr/Projects/dnapol/NJ-polB.htm). However, this RNA/DNA priming subfamily of DNA polymerase B phylogeny appeared poorly resolved. Moreover, considering that this tree is unrooted, it doesn't testify inevitably a close phylogenetic relationship. Especially, it is not possible to exclude a misplacement of the viral sequences; their long branches (relative to the Eukaryotic polymerase δ branch) being attracted by those of the other sequences that are still longer. Accordingly, we cannot rule out the hypothesis that Eukaryotic cells transferred their DNA polymerase δ to viruses.

As in the case of DNA polymerase δ, the DNA polymerase B1 of the archaeon *Halobacterium salinarum* NRC1 also emerges from a group of viral DNA polymerases. This cellular polymerase is closely related to DNA polymerases of haloviruses HF1 and HF2 and these three polymerases clustered together with the sequences of the T4 phage family. Although the node for this grouping is not robust, it seems most likely that the host *Halobacterium* polymerase is of viral origin since the HF1/HF2/*Halobacterium sp.* cluster did not branch with other archaeal DNA polymerase of the B family.

HF1 and HF2 are closely related head-tail viruses, and their combined host-ranges cover at least four genera of Halobacteria (*Haloferax, Halorubrum, Haloarcula, Halobacterium*). Their genomes share remarkable similarities with those of the bacteriophages T3 and T7 (Nuttall and Dyall-Smith 1995). Taken together, all these findings suggest a long evolutionary history for tailed phages (Ackermann 1998) rather than a recent origin followed by multiple horizontal gene transfers between bacterial and archaeal phages. Subsequent phylogenetic analysis of the other ORFs of HF1/HF2 would probably bring some valuable arguments to explain the similarity between T4 and HF1/HF2.

Conclusion

The history of the different families of DNA polymerase seem to be marked by numerous events of gene duplications, gene loss, and lateral gene transfers. Sometimes, functional shift and change in the gene environment may have lead to the rapid divergence of these sequences. These processes have made it difficult to recognize the intra-family gene homology: we cannot rule out that some very divergent molecules belonging to different families have a common evolutionary origin. In this study, we proposed that some cellular DNA polymerases could have originated from viruses and plasmids. Two likely examples are the A-type DNA polymerase Gamma of the mitochondrion, which could come from a T3/T7 related phage (as the RNA polymerase of the mitochondrion), and the B-type DNA polymerase B1 of *Halobacterium sp.* that could result from the transfer of the DNA polymerase gene of an halovirus. Taken together, these findings suggest a complex evolutionary history of the DNA replication apparatus that involved significant exchanges between viruses, plasmids, and their hosts.

Note Added at Proof

After submission of our manuscript, the name family Y was independently proposed for the group of DNA polymerases related to RAD30, UmuC, and DinB [Ohmore et al. (2001) The Y-family of DNA polymerases. *Mol Cell* 1:7–8]. We also became aware of a paper suggesting that the domain with unknown function associated to some archaeal and bacterial DNA polymerases of the X family could be a phosphoesterase domain [Aravind L, Koonin EV (1998) *Nucleic Acids Res* 26:3746–3752].

Acknowledgments. We are grateful to Dr. Henry Krisch for providing the DNA polymerase sequence of the coliphage RB49. We thank Dr. Hervé Philippe, Dr. Henry Krisch, and Dr. Michael Dyall-Smith for helpful discussions and critical reading of the manuscript. This work was funded by the PRFMMIP (Programme of the French Ministére de la Recherche).

References

Ackermann HW (1998) Tailed bacteriophages: the order caudovirales. Adv Virus Res 51:135–201

Adachi J, Hasegawa M (1996) MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Comput Sci Monogr 28:1–150

Altschul S, Gish W, et al. (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Altschul S, Koonin E (1998) Iterated profile searches with PSI-BLAST. A tool for discovery in protein databases. Trends Biochem Sci 23:444–447

Aravind L, Koonin EV (1999) DNA polymerase beta-like nucleotidyltransferase superfamily: identification of three new families, classification and evolutionary history. Nucleic Acids Res 27:1609–1618

Braithwaite D, Ito J (1993) Compilation, alignment, and phylogenetic relationships of DNA polymerases. Nucleic Acids Res 21:787–802

Burtis K, Harris P (1997) A possible functional role for a new class of eukaryotic DNA polymerases. Curr Biol 7:R743–R744

Cann I, Ishino Y (1999) Archaeal DNA replication: identifying the pieces to solve a puzzle. Genetics 152:1249–1267

Edgell D, Doolittle W (1997) Archae and the origin(s) of DNA replication proteins. Cell 89:995–998

Edgell D, Malik S, et al. (1998) Evidence of independent gene duplications during the evolution of Archeal and Eukaryotic family B DNA polymerase. Mol Biol Evol 15:1207–1217

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Sys Zoo 27:401–410

Forterre P (1992) The DNA polymerase from the archaebacterium Pyrococcus furiosus does not testify for a specific relationship between archaebacteria and eukaryotes. Nucleic Acids Res 20:1811

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

Goodman M, Tippin B (2000) Sloppier copier DNA polymerases involved in genome repair. Curr Opin Genet Dev 10:162–168

Gray MW, Lang BF (1998) Transcription in chloroplasts and mitochondria: a tale of two polymerases. Trends Microbiol 6:1–3

Harris PV, Mazina OM, et al. (1996) Molecular cloning of Drosophila mus308, a gene involved in DNA cross-link repair with homology to prokaryotic DNA polymerase I genes. Mol Cell Biol 16:5764–5771

Huang Y-P, Ito J (1999) DNA polymerase C of the thermophilic bacterium Thermus aquaticus: Classification and phylogenetic analysis of the family C DNA polymerase. J Mol Evol 48:756–769

Hübscher U, Nasheuer H-P, et al. (2000) Eukaryotic DNA polymerases, a growing family. Trends Biochem Sci 25:143–147

Ishino Y, Komori K, et al. (1998) A novel DNA polymerase family found in Archaea. J Bacteriol 180:2232–2236

Ito J, Braithwaite D (1991) Compilation and alignment of DNA polymerase sequences. Nucleic Acids Res 19:4045–4057

Johnson RE, Washington MT, et al. (1999) Bridging the gap: a family of novel DNA polymerases that replicate faulty DNA. Proc Natl Acad Sci USA 96:12224–12226

Jones DT, Taylor WR, et al. (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Koonin E, Mushegian A, et al. (1996) Non-orthologous gene displacement. Trends Genetic 12:334–336

Kornberg A, Baker T (1992) DNA replication. W.H. Freeman, New York

Leipe DD, Aravind L, et al. (1999) Did DNA replication evolve twice independently? Nucleic Acids Res 27:3389–3401

Ng WV, Kennedy SP, et al. (2000) Genome sequence of halobacterium species NRC-1. Proc Natl Acad Sci USA 97:12176–12181

Nuttall SD, Dyall-Smith ML (1995) Halophage HF2: genome organization and replication strategy. J Virol 69:2322–2327

Philippe H (1993) MUST, a computer package of management utilities for sequences and trees. Nucleic Acids Res 21:5264–5272

Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Opin Genet Dev 8:616–623

Smith DR, Doucette-Stamm LA, et al. (1997) Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics. J Bacteriol 179:7135–7155

Steitz TA (1999) DNA polymerases: structural diversity and common mechanisms. J Biol Chem 274:17395–17398

Swofford DL (1993) PAUP: phylogenetic analysis using parsimony, version 3.1.1. Illinois Natural History Survey, Champaign, IL

Thompson JD, Higgins DG, et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Villarreal LP (1999) DNA virus contribution to host evolution. In: Origin and evolution of viruses. Academic Press, San Diego, pp 391–419

Villarreal LP, DeFilippis VR (2000) A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. J Virol 74:7079–7084

Zhu W, Ito J (1994) Family A and family B DNA polymerases are structurally related: evolutionary implications. Nucleic Acids Res 22:5177–5183