Available online at www.sciencedirect.com
SCIENCE DIRECT®
Research in Microbiology 154 (2003) 237–243
Research in Microbiology
www.elsevier.com/locate/resmic

The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies

Jonathan Filée, Patrick Forterre, Jacqueline Laurent *

Laboratoire de Biologie moléculaire du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, Bât. 409, CNRS UMR 8621, Université Paris-Sud XI, 91405 Orsay Cedex, France

Abstract

Viruses are often considered as fragments of cellular RNA or DNA that escaped a long time ago from cellular chromosomes and that evolved later on by capturing additional genes from the genomes of their hosts. However, this view has now been challenged by the discovery of surprising homology between viruses with very distantly related hosts, and by phylogenetic analyses suggesting that genes might also have flown from viruses to cells. We present here phylogenetic analyses of four proteins involved in DNA replication and synthesis of DNA precursors (DNA polymerases delta, ribonucleotide reductases, thymidylate synthases and replicative helicases) and we discuss the reciprocal roles of cells and viruses during the evolutionary history of these enzymes. These analyses revealed numerous lateral gene transfer events between cells and viruses, in both directions. We suggest that lateral gene transfers from viruses to cells and nonorthologous gene replacements of cellular genes by viral ones are an important source of "genetic novelties" in the evolution of cellular lineages. Thus, viruses have definitively to be considered as major players in the evolution of cellular genomes.

Keywords: Nonorthologous gene displacement; Horizontal gene transfer; Phylogeny; Virus; Phage; Plasmid; Mitochondria

1. Introduction

Despite their huge biodiversity, viruses and other mobile genetic elements, like plasmids or transposons, remain considerably understudied by evolutionary biologists, in comparison to their cellular hosts. Their potential role in host evolution is often underestimated and viral genes are rarely used in molecular phylogeny studies. The main explanation for this situation is that viruses are often conceived as deriving mostly from cellular DNA, which recently became autonomous. Accumulating evidence, however, supports the view that viruses probably diverged before the LUCA (the last universal cellular ancestor). A major argument in favour of this hypothesis is the existence of numerous striking structural similarities between viruses that infect organisms belonging to the different domains of life. For instance, archaeal and bacterial tailed phages show remarkable morphological similarity [24] and the archaeal viruses SIRV show significant structural and mechanistic similarities to eukaryal Poxviruses [17]. Moreover, some eukaryal and bacterial viruses exhibit a high level of similarity with regard to the organisation of their genomes and replication process [5]. Finally, studies on the structure of capsid proteins have provided evidence for relatively close relationships between some bacteriophages and eukaryal viruses [10].

This long evolutionary history of viruses opens up interesting questions about the role played by viral and cellular forms of life in their respective evolutions. As viruses with large and double-stranded DNA genomes have many genes homologous to cellular genes, they are generally thought to have evolved by capturing multiple genes from their hosts. In this case, in a phylogenetic tree, viral sequences are likely to be sister groups of the corresponding sequences from their cellular host. A good illustration of this is given by Moreira [15] who showed that, for three different bacterial proteins (replicative helicase, Ssb protein and topoisomerase III), some phage and plasmids sequences could be related to hosts sequences through multiple horizontal gene transfers from cells to phages. Actually, a still more complex situation is conceivable, and probably more appropriate, since extant genomes might be mosaics of sequences of various origins, resulting from horizontal gene transfers from cells to viruses, but also from viruses to cell [11,22,24] or from nonorthologous gene displacements [4]. Nonortholo-

* Corresponding author.
E-mail address: jacqueline.laurent@igmors.u-psud.fr (J. Laurent).

0923-2508/03/$ – see front matter
doi:10.1016/S0923-2508(03)00066-4

gous gene displacement refers to situations where unrelated, or paralogous proteins are responsible for the same critical functions in different species [13]. Forterre suggested that plasmids and viruses were the donors of nonorthologous genes that replaced ancestral cellular versions in the evolution of the DNA replication mechanisms [6]. The RNA and DNA polymerases of the mitochondrion appear to be good examples of such nonorthologous displacement of a cellular gene by a viral version. Indeed, phylogenetic analyses indicate that they are more related to bacteriophage T3/T7 RNA polymerase [8] and T3/T7 DNA polymerase, than to RNA and DNA polymerases from their bacterial ancestor [4].

In this work, we were interested in further documenting the reciprocal role of gene exchange between viruses and their hosts in their respective evolution. We focused our work on enzymes involved in DNA metabolism and replication because, in addition to cellular sequences, numerous viral and plasmid sequences are available in the databanks for these proteins. Based on several phylogenetic studies, our work confirms that, as previously suggested, lateral gene transfer between viruses and their hosts seem to be a major factor in viral genome evolution.

2. Methods

For each family of proteins that we have studied, a representative query sequence was chosen as the query for a BLAST search [2]. The homologous sequences were retrieved using the program ALIBABA (Philippe Lopez, personal communication) and aligned with each other using CLUSTAL W [21]. The alignment was refined manually with the help of the ED program of the MUST package version 3.0 [18]. Positions that could not be unambiguously aligned were excluded from the analysis and gaps were removed.

Phylogenetic trees were constructed with maximum likelihood (ML) and distance-based methods, using the programs PROTML [1] (version 2.3) and NJ in the MUST package [18] (version 3.0), respectively. ML trees were obtained by a quick-add search with 1000 replications using the JTT-F model of substitution and bootstrap values were calculated using the RELL method with the BOOTML program (Philippe, personal communication). 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 (version 3.0) [18].

3. Results

3.1. Phylogeny of DNA polymerases

We previously performed an extensive phylogenetic analysis of the five main DNA polymerase families (A, B, C, X, and Y) [4]. Viral and plasmid genes encoding for their own DNA polymerases are widespread in several of these families (A, B, Y and to a lesser extent C). This analysis has shown that multiple gene exchanges occurred between viruses, plasmids and hosts in almost all these families. In particular, we obtained evidence that the ancestral replicase of the mitochondria or their bacterial ancestor (from family C) was replaced by a nonhomologous DNA polymerase of viral origin (related to bacteriophage T3/T7) that became the mitochondrial DNA polymerase gamma [4].

In agreement with a previous report by Villarreal [23], our global phylogenetic analysis of family B also suggested that eukaryal DNA polymerase delta was of viral origin [23]. Here we performed a phylogenetic analysis of the B DNA polymerases of the subclass "delta", rooted with the zeta subclass to obtain more precise information on this issue (Fig. 1). In addition to cellular DNA polymerases, the subclass delta includes genes belonging to several groups of eukaryotic viruses: Herpesvirus, Phycodnavirus, Ascovirus and Iridovirus. By focusing our analysis, we could use a larger number of positions (278 amino acids). Moreover we constructed the tree using a maximum likelyhood approach, which is not very sensitive to the difference in evolutionary rates among sequences that can otherwise dramatically affect phylogenetic trees in which viral and cellular sequences are mixed [15]. The resulting tree is presented in Fig. 1. The monophyly of the cellular sequences, on the one hand, and of the viral sequences, on the other hand, are strongly supported, but viral genes occupy the base of the tree with a low statistical support. These results might indicate that an ancient gene exchange indeed occurred between an ancestral virus and its host before the radiation of eukaryotic cellular lineages. However, the polarity of this transfer cannot be determined.

3.2. Phylogeny of ribonucleotide reductases

Ribonucleotide reductases (RNRs) are key enzymes in the transition from an "RNA world" to a "DNA world". They have been classified into three classes (I, II and III) according to their subunit composition and cofactor requirement. Class I is present in Bacteria and Eukarya, whereas class II and III are present in Bacteria and Archaea. Many viruses also encode their own RNRs. Mechanistic and structural similarities indicate that all RNRs are built around a homologous catalytic core subunit [19]. However, extensive sequence similarities can only be detected between class I and II catalytic subunits [12]. Here we present the phylogeny of the homologous subunits of families I and II. In order to deal with a larger number of amino acid positions, only prokaryotic sequences were used to build up the NJ tree presented in Fig. 2.

In this analysis, the RNR of the bacteriophage SPBC2 and that of the archaeal phage HF2 are closely related to the RNR of their hosts (*Bacillus subtilis* and *Halobacterium* sp., respectively). The RNR of *B. subtilis* and of the

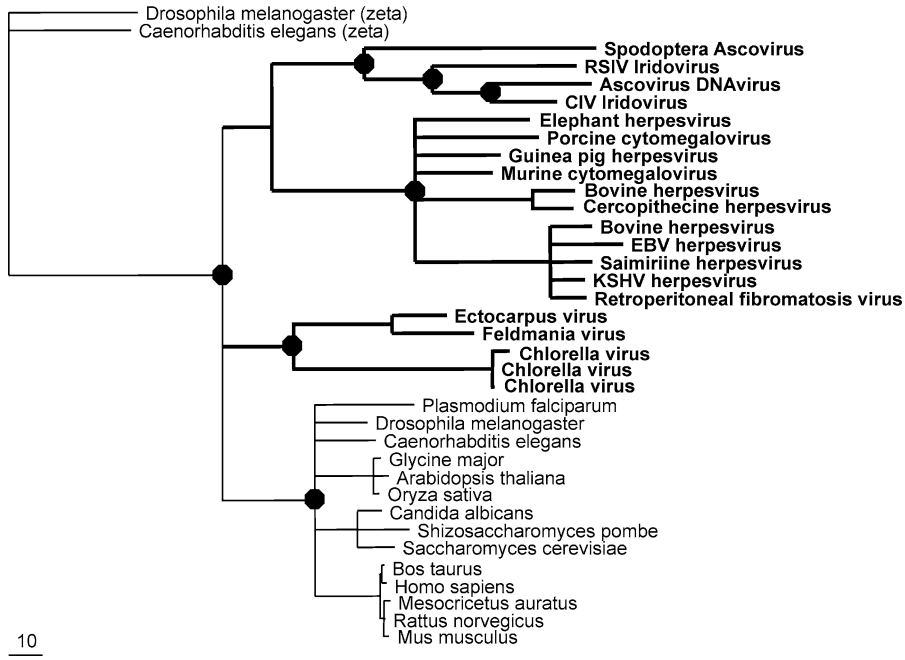

Fig. 1. Phylogenetic tree of B-type DNA polymerase belonging to the "Delta" family. The tree was rooted using the sequences of the "Zeta" family. Bootstrap values higher than 90% are indicated by a filled circle, and those lower than 50% were collapsed. Viral sequences are indicated in bold. The scale bar represents the number of substitutions per 100 sites for unit branch length.

bacteriophage SPBC2 are clustered together with RNR of other Gram-positive bacteria, while the RNR of HF2 and *Halobacterium* are clustered with other archaeal RNR. This suggests that these viruses have acquired their RNR from their hosts rather than the reverse. The sequence of the coliphage T4 RNR also appeared to be closely related to one of the RNRs from its host (*E. coli*) (called type Ia, 12) in a cluster of RNRs from Proteobacteria. However in that case, the situation is more complex, since *E. coli* and another member of this cluster, *Salmonella typhimurium*, possess another RNR (Ib) which is located in a large group of RNRs from various bacterial phyla. This might indicate that the Ib RNR corresponds to the ancestral RNR present in Bacteria, whereas the "Ia" RNR of Proteobacteria has a viral origin. In this hypothesis, the two RNRs were conserved in *S. typhimurium* and *E. coli*, whereas one of them was selectively lost in the other Proteobacteria.

3.3. Phylogeny of thymidylate synthase (TdS)

TdS have been essential in the transition from the DNA-U world (DNA containing uracil) to a DNA-T world (DNA containing thymidine) since these enzymes are required for the synthesis of dTMP by methylation of dUMP. For a long time, all TdS were believed to be homologous. However, it turned out recently that two families of nonhomologous TdS, ThyA and ThyX are present in living organisms [16]. Numerous viral sequences are available in the databanks for these two proteins. Here we present the ML phylogeny of the ThyA protein (Fig. 3), whereas we published elsewhere the phylogeny of ThyX [16]. In the eukaryotic ThyA tree, several viral lineages are dispersed between various eukaryotic kingdoms, whereas a series of herpes viruses are grouped together with the mammalian sequences. The later grouping argues for the acquisition of these herpes viral TdS from their hosts. In contrast, the erratic dispersion of the other viral TdS in the eukaryotic tree renders uncertain any conclusion about their phylogenetic positions.

Interestingly, all prokaryotic ThyA were grouped into two clusters in our phylogeny, one that includes only Bacteria, and the other that includes Bacteria, bacteriophages and one archaeon (*Haloferax volcani*). In this second group, the viral and cellular ThyA proteins are mixed up. Some viral proteins, such as those from the *Bacillus subtilis* bacteriophages beta-22 and phi-3T are located close to the ThyA proteins of their hosts, whereas others, such as ThyA of bacteriophages T4 and pHCM2 are located at the base of the group. Some bacteria, for exemple *B. subtilis*, have two ThyA proteins, one from each group. This suggests a viral origin for the cellular ThyA proteins included in the group containing both cells and viruses. However, the long branches of all members of this group in the ThyA tree represent a powerful source of bias, so that their grouping might also result from the long branch attraction (LBA) artefact, rather than reflecting evolution from an actual common ancestor gene. Consequently, our conclusion about the viral origin of this cluster has to be taken with caution.

Interestingly, ThyA and ThyX have sporadic phylogenetic distributions and this repartition is mutually exclusive [16]; when ThyX is present, ThyA is absent, and vice versa (except for *Mycobacterium* species, whose genomes harbour the two classes). Moreover, as shown here for the

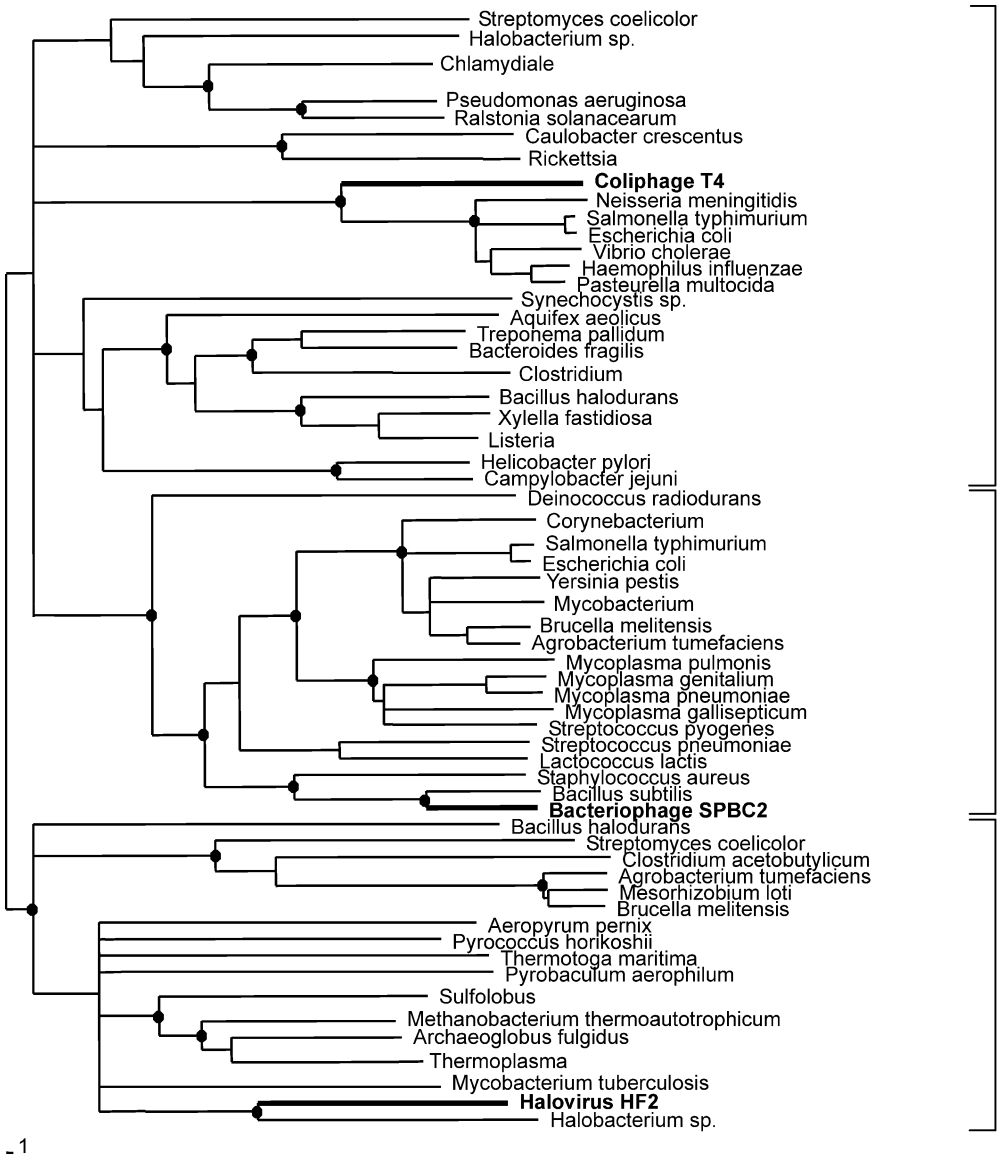

Fig. 2. Unrooted phylogenetic tree of the prokaryotic families I and II of the ribonucleotide reductase. Bootstrap values higher than 90% are indicated by a filled circle, and those lower than 50% were collapsed. Viral sequences are indicated in bold. The scale bar represents the number of substitutions per 100 sites for unit branch length.

gene *thyA*, numerous events in lateral gene transfer between viruses and cells have also affected the gene *thyX*. These observations are in agreement with the idea that many independent nonorthologous gene displacements of one class of TdS by the other occurred in various cellular lineages.

3.4. Phylogeny of prokaryotic and mitochondrial replicative helicases

Some bacteriophages encode a helicase-primase protein whose helicase domain is homologous to the bacterial helicase involved in DNA replication: DnaB. A phylogeny of these proteins was previously published by Moreira (2000). His analysis led to the conclusion that several *dnaB* genes were transferred from Bacteria to bacteriophages. More recently, it was noticed in silico [14] and demonstrated experimentally [20] that mammals contain a DnaB-like helicase which is involved in mitochondrial DNA replication. We thus decided to perform a new phylogenetic analysis of the DnaB family. BLAST search seeded with the mitochondrial DnaB sequence of *Homo sapiens* retrieved with highly significant E value several other eukaryotic sequences (Metazoa, and *Plasmodium*, E value $<10^{-50}$, and *Arabidopsis*, E value $= 10^{-7}$) and several sequences of bacteriophages belonging to the T3/T7 group (E values between 10^{-9} and 10^{-5}). Canonical bacterial DnaB sequences were retrieved with less significant E values, starting with 10^{-3}. The ML phylogeny of these sequences rooted with the bacterial DnaB is presented in Fig. 4. Eukaryotic sequences are a sister group of the T3/T7 phage sequences with high sta-

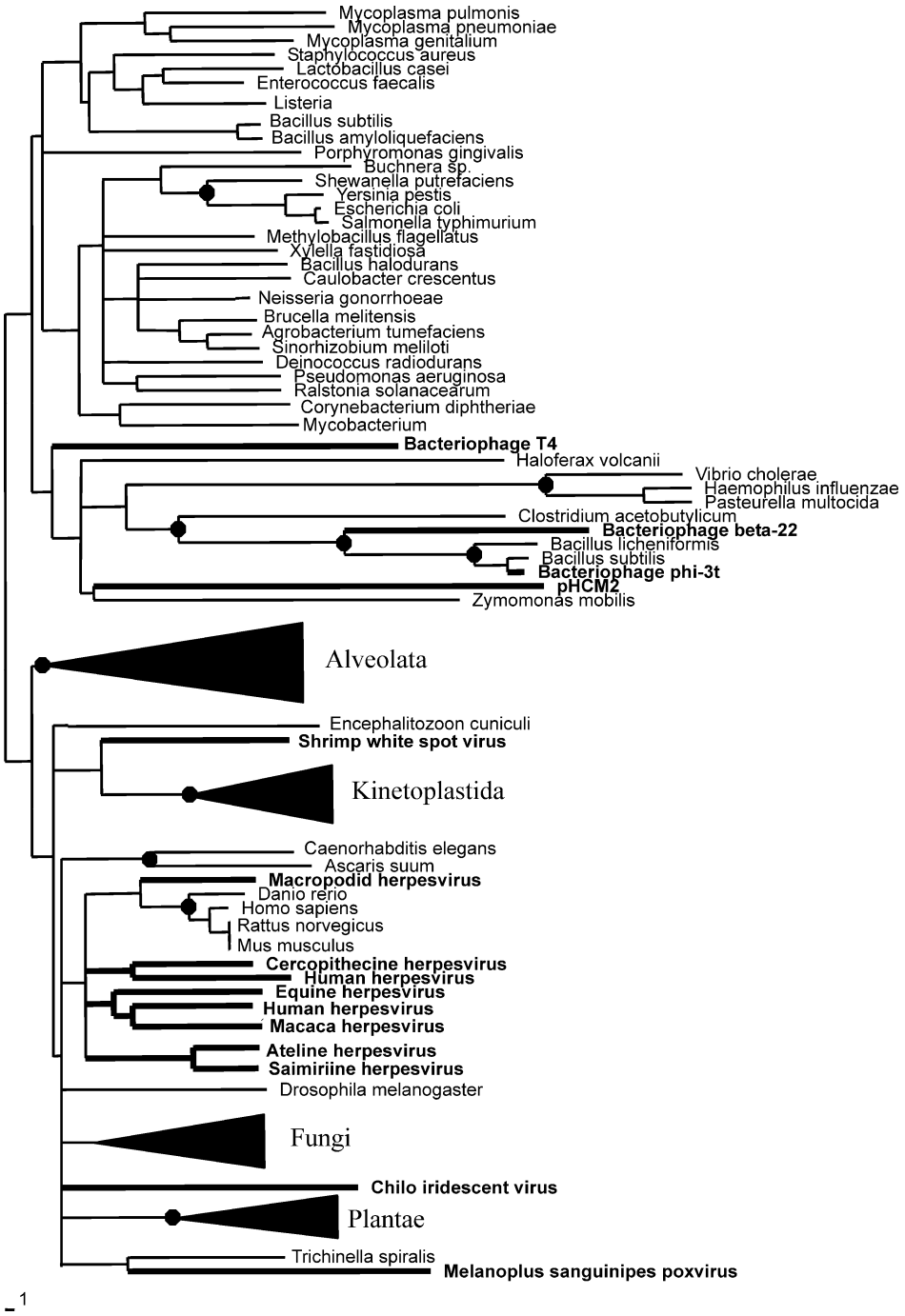

Fig. 3. Unrooted phylogenetic tree of thymidylate synthase (ThyA type). Bootstrap values higher than 90% are indicated by filled circle, and those lower than 50% were collapsed. Viral sequences are indicated in bold. The scale bar represents the number of substitutions per 100 sites for unit branch length.

tistical support. As mitochondria derived from a proteobacterium, their mitochondrial DnaB should have formed a monophyletic group with bacterial ones if they were derived from the helicase of the endosymbiont. Our result indicates instead that the mitochondrial DnaB most likely originated from a T3/T7 related phage rather than from a bacterium. This is reminiscent of the situation reaveled by mitochondrial RNA polymerase [9] and DNA polymerase [4] phylogenies. All these data suggest that a global event in horizontal gene transfer involving at least three genes occurred between mitochondria (or their bacterial ancestor) and a phage belonging to the T3/T7 group.

4. Discussion

This study, in analysing the phylogenies of several proteins implicated in the replication and the metabolism of

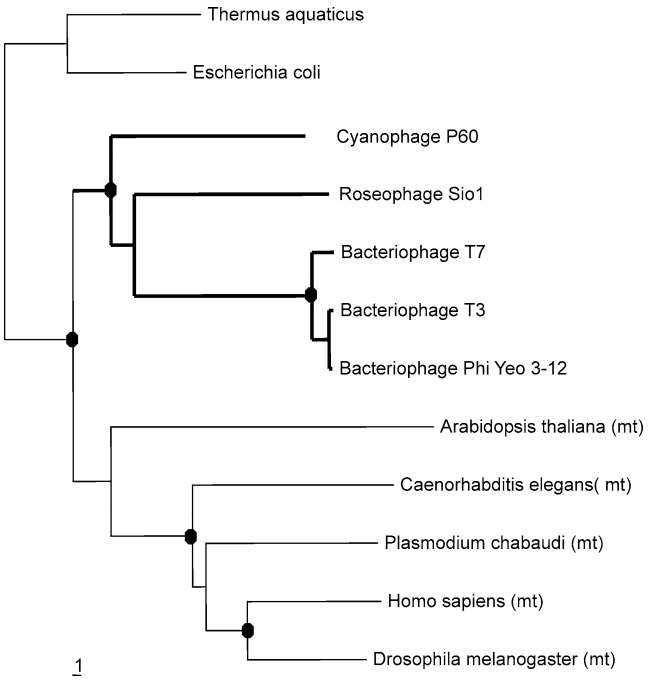

Fig. 4. Phylogenetic tree of the replicative helicase DnaB rooted with the canonical bacterial DnaB. Bootstrap values higher than 90% are indicated by filled circle, and those lower than 50% were collapsed. Viral sequences are indicated in bold. The scale bar represents the number of substitutions per 100 sites for unit branch length.

DNA, revealed various evolutionary relationships between viruses and their hosts. We can principally identify three major types of relationships.

4.1. Horizontal gene transfer from cells to viruses

In our study, horizontal gene transfer from cells to viruses was strongly suggested for the ribonucleotide reductase of the archaeal phage HF2 and the *Bacillus* phage SpBc. This case is well known, and many other examples exist in the literature, such as the UTPase of the *Bacillus* phage SpBc [3], or the single strand binding protein (Ssb) of the coliphage P1 [15]. As the mutational rates of viral proteins are thought to be higher than those of their hosts, interesting questions are raised about the destiny of these cellular genes in their new environment. Many of them are probably lost in the course of evolution and could transiently become pseudogenes. But others could be conserved if they provided any selective advantage to the virus and could possibly acquire new properties (such as high processivity in the case of DNA polymerases).

4.2. Horizontal gene transfer from viruses to cells

When several copies of the same gene are present in a cellular genome, and one of them is closely related to a viral counterpart, the simplest explanation is that the latter gene has a viral origin. In our study, this is the case for the Ia class of the ribonucleotide reductase of Proteobacteria, and for one of the two copies of the TdS (ThyA) of the *Bacillus* species. A similar situation exists for the DNA polymerase delta of eukaryotes [23] and for the DNA polymerase B1 of *Halobacterium* sp. [4]. The alternative explanation for such a situation implies the duplication of the cellular gene, then followed by the acquisition of one paralogue by a virus. But this explanation does not account for the phylogenetic pattern observed in the examples quoted above. Indeed, the cellular proteins that are suspected to be of viral origin are generally located far from the other cellular proteins in the tree (e.g., the cluster grouping ThyA of *Bacillus* species with bacteriophage Phi-3t and beta-22 ThyA is located far from all the other Gram-positive bacteria, Fig. 2).

4.3. Nonorthologous gene displacement of cellular gene by a viral gene

This is a special case of the horizontal gene transfer polarised from viruses to cells, i.e., when the viral gene replaced a nonorthologous cellular version. Two examples are documented in the literature for mitochondrial RNA and DNA polymerases which both came from a T3/T7 type bacteriophage [4,8]. Here we show that the replicative helicase DnaB of the mitochondrion also probably originated from a T3/T7 type bacteriophage. In addition, the existence of two different TdS, and the respective phylogeny of each protein ([16] for ThyX and this study for ThyA) testify to the existence of multiple and independent events of nonorthologous gene displacement of a cellular gene by a viral one. This is probably the case for the thymidylate synthase ThyX of *Mycobacterium species*, which are closely related to that of the temperate Mycobacteriophages D29 and L5. Such acquisition by a cellular host of a viral gene with no clear homolog in its genome could be a powerful source of genetic novelty. Finally, this has raised the question of the possible invention of a new gene by a virus followed by a transfer into the genome of its host.

The integration of cryptic prophages which have lost their ability to excise and replicate themselves seems to be a common mechanism to acquire viral genes. For example, one of the C-type DNA polymerases of *B. subtilis* is located in a cryptic prophage genome [4]. Temperate bacteriophages which integrate their genomes in the host genomes, or that exist in a carrier state for a long evolutionary period, could also be a source of new cellular genes. The most likely process for these captures of viral genes by cellular genomes is illegitimate recombination events mediated by transposons or other mobile elements

The existence of several nonhomologous genes for accomplishing the same biological function is well documented [7]. This situation is crucial for the DNA replication apparatus, which is very different in Bacteria compared to Archaea and Eukarya. It was suggested that viruses have played a central role in this phenomenon [4,6,22]. Our results are in agreement with these ideas, since we were able to show that several cellular genes, especially in bacteria, may

have a viral origin. However, much more data are needed to substantiate these scenarios. Several phylogenies are difficult to interpret because of a high level of gene exchange between viruses and host cells, the loss of a phylogenetic signal for ancient phylogenies, and the restricted sampling of viral sequences. Hopefully, the systematic sequencing of new viral genomes from different virus families and from viruses infecting very divergent species will in the end provide decisive data for better understanding the evolution of viruses and for confirming the importance of their contribution to the evolution of their hosts.

Acknowledgements

This work was supported by a Grant from the Programme de Recherche Fondamentale en Microbiologie et Maladie Infectieuse et Parasitaires du Ministère National de l'Education et de la Recherche (MNER). Jonathan Filée was supported by a fellowship from the MNER.

References

[1] J. Adachi, M. Hasegawa, MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood, Comput. Sci. Monogr. 28 (1996) 1–150.
[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[3] A.M. Baldo, M.A. McClure, Evolution and horizontal transfer of dUTPase-encoding genes in viruses and their hosts, J. Virol. 73 (1999) 7710–7721.
[4] J. Filée, P. Forterre, T. Sen-Lin, J. Laurent, Evolution of DNA polymerase families: Evidences for multiple gene exchange between cellular and viral proteins, J. Mol. Evol. 54 (2002) 763–773.
[5] P. Forterre, The DNA polymerase from the archaebacterium *Pyrococcus furiosus* does not testify for a specific relationship between archaebacteria and eukaryotes, Nucleic Acids Res. 20 (1992) 1811.
[6] P. Forterre, Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins, Mol. Microbiol. 33 (1999) 457–465.
[7] M.Y. Galperin, E.V. Koonin, Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes, Genetica 106 (1999) 159–170.
[8] M. Gray, B. Lang, Transcription in chloroplasts and mitochondria: A tale of two polymerases, Trends Microbiol. 6 (1998) 1–3.
[9] B. Hedtke, T. Borner, A. Weihe, Mitochondrial and chloroplast phage-type RNA polymerases in *Arabidopsis*, Science 277 (1997) 809–811.
[10] R.W. Hendrix, Evolution: The long evolutionary reach of viruses, Curr. Biol. 9 (1999) R914–917.
[11] A.L. Hughes, Origin and evolution of viral interleukin-10 and other DNA virus genes with vertebrate homologues, J. Mol. Evol. 54 (2002) 90–101.
[12] A. Jordan, P. Reichard, Ribonucleotide reductases, Annu. Rev. Biochem. (1998) 71–98.
[13] E.V. Koonin, A.R. Mushegian, P. Bork, Nonorthologous gene displacement, Trends Genet. 12 (1996) 334–336.
[14] D.D. Leipe, L. Aravind, N.V. Grishin, E.V. Koonin, The bacterial replicative helicase DnaB evolved from a RecA duplication, Genome Res. 10 (2000) 5–16.
[15] D. Moreira, Horizontal transfer of informational genes, Mol. Microbiol. 35 (2000) 1–5.
[16] H. Myllykallio, G. Lipowski, D. Leduc, J. Filée, P. Forterre, U. Liebl, An alternative flavin-dependent mechanism for thymidylate synthesis, Science 297 (2002) 105–107.
[17] X. Peng, H. Blum, Q. She, S. Mallok, K. Brugger, R.A. Garrett, et al., Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: Relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses, Virology 291 (2001) 226–234.
[18] H. Philippe, MUST, a computer package of Management Utilities for Sequences and Trees, Nucleic Acids Res. 21 (1993) 5264–5272.
[19] M.D. Sintchak, G. Arjara, B.A. Kellogg, J. Stubbe, C.L. Drennan, The crystal structure of class II ribonucleotide reductase reveals how an allosterically regulated monomer mimics a dimer, Nat. Struct. Biol. 9 (2002) 293–300.
[20] J.N. Spelbrink, F.Y. Li, V. Tiranti, K. Nikali, Q.P. Yuan, M. Tariq, et al., Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria, Nat. Genet. 28 (2001) 223–231.
[21] J.D. Thompson, D.G. Higgins, T.J. Gibson, 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 (1994) 4673–4680.
[22] L.P. Villarreal, in: R.W.E. Domingo, J.J. Holland (Eds.), DNA Virus Contribution to Host Evolution, Academic Press, San Diego, 1999.
[23] L.P. Villarreal, V.R. DeFilippis, A hypothesis for DNA viruses as the origin of eukaryotic replication proteins, J. Virol. 74 (2000) 7079–7084.
[24] W. Zillig, D. Prangishvilli, C. Schleper, M. Elferink, I. Holz, S. Albers, et al., Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea, FEMS Microbiol. Rev. 18 (1996) 225–236.