Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochimie 87 (2005) 793–803
BIOCHIMIE
www.elsevier.com/locate/biochi

The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells

Patrick Forterre [a,b,*]

[a] *Institut Pasteur, Département de Microbiologie Fondamentale et Médicale, 25, rue du Docteur Roux, 75015 Paris, France*
[b] *University Paris-Sud, Institut de Génétique et Microbiologie, CNRS, UMR8621, 91405 Orsay cedex, France*

Received 3 January 2005; accepted 18 March 2005

Available online 12 April 2005

Abstract

Most evolutionists agree to consider that our present RNA/DNA/protein world has originated from a simpler world in which RNA played both the role of catalyst and genetic material. Recent findings from structural studies and comparative genomics now allow to get a clearer picture of this transition. These data suggest that evolution occurred in several steps, first from an RNA to an RNA/protein world (defining two ages of the RNA world) and finally to the present world based on DNA. The DNA world itself probably originated in two steps, first the U-DNA world, following the invention of ribonucleotide reductase, and later on the T-DNA world, with the independent invention of at least two thymidylate synthases. Recently, several authors have suggested that evolution from the RNA world up to the Last Universal Cellular Ancestor (LUCA) could have occurred before the invention of cells. On the contrary, I argue here that evolution of the RNA world taken place in a framework of competing cells and viruses (preys, predators and symbionts). I focus on the RNA-to-DNA transition and expand my previous hypothesis that viruses played a critical role in the emergence of DNA. The hypothesis that DNA and associated mechanisms (replication, repair, recombination) first evolved and diversified in a world of DNA viruses infecting RNA cells readily explains the existence of viral-encoded DNA transaction proteins without cellular homologues. It also potentially explains puzzling observations from comparative genomic, such as the existence of two non-homologous DNA replication machineries in the cellular world. I suggest here a specific scenario for the transfer of DNA from viruses to cells and briefly explore the intriguing possibility that several independent transfers of this kind produced the two cell types (prokaryote/eukaryote) and the three cellular domains presently known (Archaea, Bacteria and Eukarya).

Keywords: RNA world; RNA/DNA transition; Virus origin; LUCA; universal tree of life

1. Introduction

All present-day cellular organisms have DNA genomes. The origin of DNA is thus a central issue for people interested in the first steps of life history. There is now a consensus among evolutionists that DNA genomes were preceded by RNA genomes (for recent reviews see Ref. [1]). This is reasonable considering, among other things, that DNA can be viewed as a modified form of RNA. Ribose is the normal sugar (not deoxyribose) and thymidine is 5-methyl-uracile. However, there is no consensus about the nature of the RNA world and the very meaning of this term. I will start this paper by a brief description of my own view and a definition of this fascinating period of early life evolution, with emphasis on a proposal to distinguish two periods (two ages) in the RNA world.

* Tel.: +33 1 69 15 74 89; fax: +33 1 69 15 78 08.
E-mail address: forterre@pasteur.fr (P. Forterre).

0300-9084/$ - see front matter © 2005 Elsevier SAS. All rights reserved.
doi:10.1016/j.biochi.2005.03.015

2. The RNA world as a cellular world

Life is a concept; in the material world, there are only living organisms (open thermodynamic systems in competition with each others). Speaking about the RNA world, one should then first ask what kinds of organisms were living there? The term RNA world was first coined by Gilbert [2] to emphasise a world of free-living RNA molecules in competition with each others. In such definition, either RNA molecules are considered as «living molecules» or the RNA world should be considered itself as «prebiotic». In fact, it seems to me (and others) very unlikely (if not impossible) that a world of free molecules could have evolved to such an extent to produce a set of complex ribozymes able to synthesise proteins (see for Ref. [3]). This is not a universally accepted assumption; indeed, until recently, several authors have suggested that cellularisation only occurred late in life history, i.e. after the divergence of the three domains of life (Archaea, Bacteria

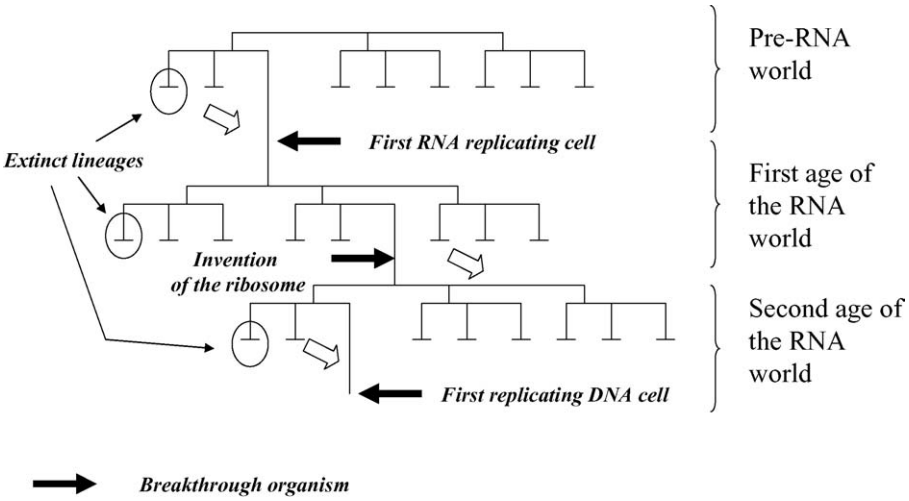

Fig. 1. The two ages of the RNA world.
Evolution goes through several critical steps each characterised by a bottleneck (a breakthrough organism). During each step, many lineages diverge from the previous breakthrough organisms. All of them but one became extinct during the transition from one step to the others. Some cellular lineages could possibly survived as viral lineages if they parasite the successful one (white arrow, and see Fig. 2).

and Eukarya) [4–6]. In particular, these authors have suggested that only a late emergence of membranes could explain why archaeal lipids are so different from eucaryotic/bacterial lipids (with opposite stereochemistry and different backbones for the long carbon chains). However, this hypothesis, which implies an acellular Last Universal Common Ancestor (LUCA), is contradicted by phylogenomics analyses showing that several membrane-related proteins were already present in LUCA. This is the case for some enzymes involved in lipid biosynthesis, for the signal recognition particle, and for the V/F-ATPases (for recent review, see Ref. [7]). These observations can be considered as definitive arguments that LUCA was a cellular organism. In fact, cellularisation most likely arose much earlier [3]. It seems difficult (if not impossible) to imagine the early development of an elaborated metabolism in the absence of cellular confinement. Let's remind that such early metabolism in the RNA world should have been able to produce at least precursors for RNA and lipid syntheses, as well as the associated energy production required to perform these reactions. Accordingly, I will define here the RNA world as a biosphere of cells with RNA genomes (RNA cells). This RNA world started with the first RNA-cell and was over when all its cellular descendants were eliminated in the Darwinian competition by cells with DNA genomes (DNA-cells).

3. The two ages of the RNA world and the "invention" of modern proteins

It is now clear that the formation of the peptide bond in modern cells is catalysed by the RNA moiety of the ribosome [8]. A crucial transitional step in the history of the RNA world was therefore the emergence of the ribozyme ancestor of today's ribosomes, together with the establishment of an earliest version of the present genetic code, i.e. the "invention" of modern proteins by RNA. In order to clarify discussions about early life evolution, I suggest distinguishing the two periods that occurred *before* and *after* this event as the *first* and *second* age of the RNA world, respectively (Fig. 1). This has the advantage to precise what we are talking about when we speculate about the ancient RNA world. In particular, this nomenclature allows asking questions such as: did DNA appear during the first or the second age of the RNA world?

The period called here "the second age of the RNA world" is often mentioned in the literature as the ribonucleoprotein world (RNP world). However, although RNA cells obviously did not contain «modern» proteins (made in ribosome) during the first age of the RNA world, this does not mean that early RNA cells did not contain peptides or protein-like molecules synthesised by other pathways that have today disappeared [9]. The term RNP world is therefore somewhat ambiguous, except if one specify that it already involved ribosome-made proteins.

It seems logical to imagine that many competing lineages of early RNA cells independently invented in the first age different mechanisms to assemble amino acids into different types of peptides (for instance in terms of amino-acid diversity or chirality) (Fig. 2A). What we know is that all these ancient lineages of RNA cells were later on out-competed by the descendants of the RNA-cell containing the ancestor of present-day ribosomes, opening the stage for the second age of the RNA world. Similarly, during this second age, many lineages of RNA cells with modern proteins should have diversified and compete until one of them (or several see below) gave birth to the first lineage(s) of DNA-cells (Figs. 1 and 2).

4. The RNA world and the origin of viruses

In both ages of the RNA world, a variety of organisms should have coexisted with preys and predators, free-living

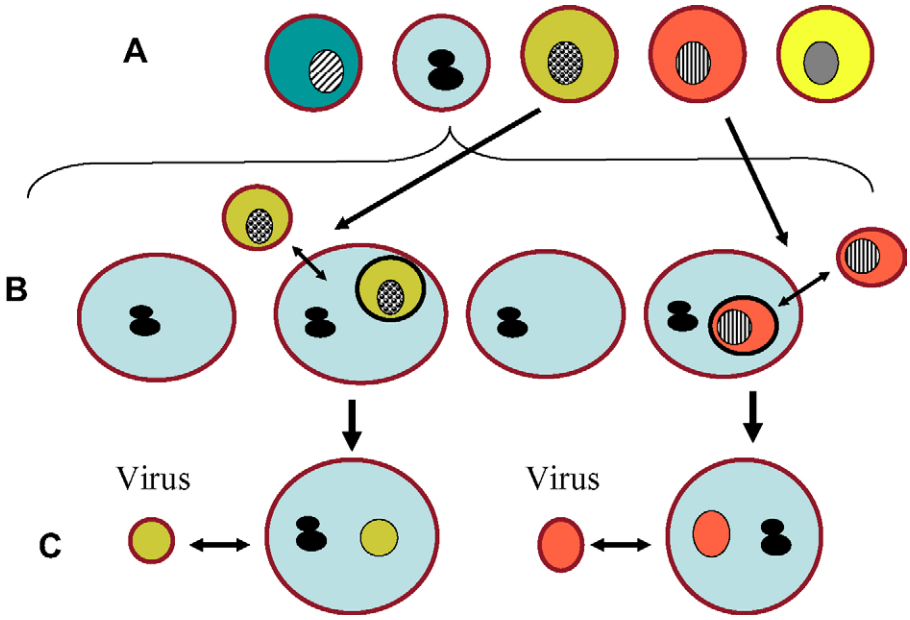

Fig. 2. Origin of RNA viruses from RNA cells.
A: Various cell lineages (different colours) coexisted in the second age of the RNA world and invented different mechanisms to produce proteins (small internal circles) including the ancestor of the present ribosome-based system (black two subunits ribosomes). B: The later lineage (in blue) eliminated all its competitors. Some of them (red and green) survived as intracellular parasites of the successful lineage with an extra-cellular stage in their life cycle. C: The parasites lose their own protein-synthesising machineries and became RNA viruses. This model implies a polyphyletic origin for different viral super-families. It can be accommodate to explain the origin of some lineage of DNA viruses in the early DNA world [10].

cells and cellular parasites. As a consequence, it is very likely that cells and virus-like organisms already coexisted and fought each other (or cohabited in various ways) in the RNA world. Since all present-day viruses contain proteins, the first viruses most likely originated as RNA viruses in the second age of the RNA world. I suggested some time ago that viruses evolved by parasitic reduction from ancient cellular lineages that were out-competed in the Darwinian selection process before LUCA, and thus could only survive as parasites in the winner of this competition [10]. In this model, RNA viruses originated from RNA cells (white arrows in Fig. 1). For instance, one can imagine an RNA-cell with a poorly efficient protein-synthesising machinery living as a parasitic endosymbiont in another RNA-cell equipped with a more efficient one (Fig. 2B). In such condition, one can easily imagine that the former may give up its poorly efficient machinery to rely completely upon that of its host, becoming a virus (Fig. 2). My vision of the RNA world is thus one of coevolving populations of very diverse RNA cells, and of RNA viruses with a variety of complex relationships and molecular mechanisms which have disappeared today, except for those which were retained in the first DNA cell or in viruses that later on infected its descendents.

5. DNA most likely originated during the second age of the RNA world

We can now go back to the previous question: did DNA appear during the first or the second age of the RNA word? In other words, could it be that the crucial transformation of RNA into DNA was initially performed by a ribozyme? In modern cells, DNA is made from RNA precursors, rNTPs, which are converted into DNA precursors, dNTPs (another argument in favour of the anteriority of RNA) (Fig. 3A). This conversion is catalysed by two types of enzymes: ribonucleotide reductases and thymidylate synthases. Ribonucleotide reductases catalyse the conversion of rNTP into dNTP (for some of them at the diphosphate stage), whereas thymidylate synthases catalyse the modification (methylation) of dUMP into dTMP. The message encoded by an RNA molecule can be also copied into DNA by viral reverse transcriptases.

It seems reasonable that the enzymes that transform RNA into DNA in modern cells are the descendants of those that presided to the historical RNA/DNA transition. The most critical reaction in this process, on a chemical point of view, is the reduction of ribose into deoxyribose by ribonucleotide reductases. Three classes of ribonucleotide reductases have been discovered up to now (named I, II and III). All of them are large elaborated proteins (using various cofactors) built around an homologous protein domain that performs a complex radical-based reaction (for a recent review, see Ref. [11]). It has been convincingly argued that this reaction could not have been be performed by an RNA molecule, especially because RNA would be too sensitive to the presence of the radical intermediates [12]. If this is true, one can safely assume that, in contrast to modern proteins, DNA was not «invented» by RNA, but that RNA was modified into DNA by protein–enzymes. According to our definition, this means that DNA appeared during the second age of the RNA world.

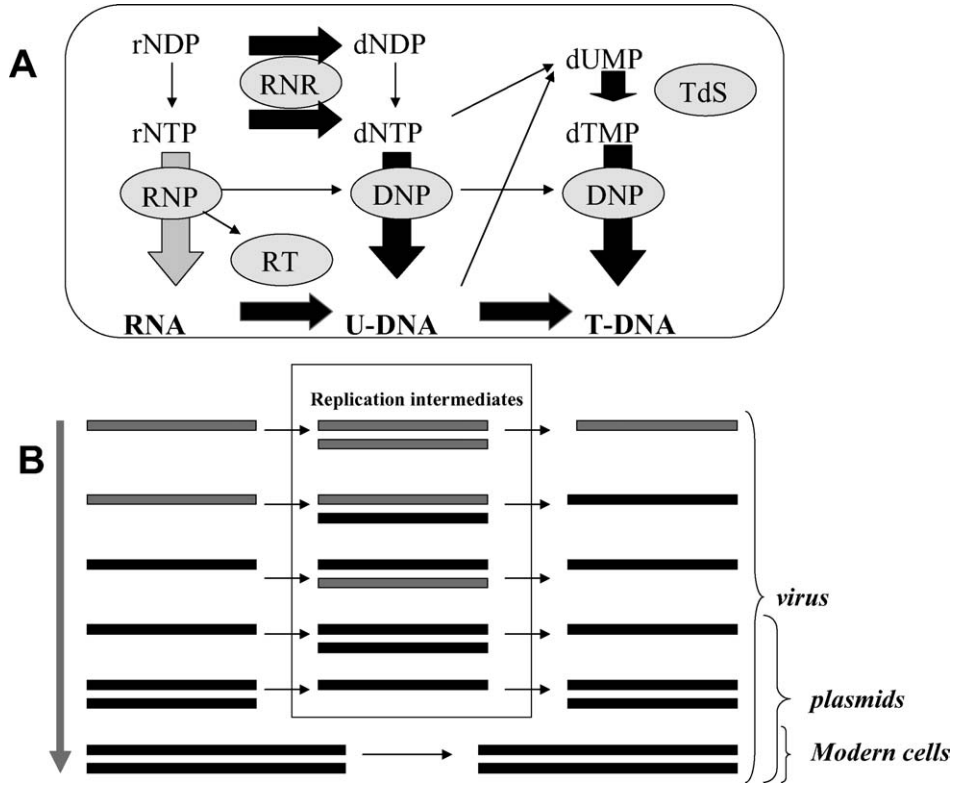

Fig. 3. Transition from RNA to DNA.
A: Enzymatic pathways required to sustain transitions from RNA to U-DNA and T-DNA. RNA: ribonucleotide reductase, TdS: thymidylate synthase. RNP: RNA polymerase, DNP: DNA polymerase, RT: reverse transcriptase. B: Evolution of genome replication mechanisms. Grey bars, RNA; black bars, DNA. The grey arrow suggests an evolutionary pathway from single-stranded RNA genomes to double-stranded DNA genomes. All mechanisms are observed in present-day viruses, whereas most plasmids have DNA genomes and all cells have double-stranded DNA genomes.

6. Complex RNA cells might have existed in the second age of the RNA world

The above scenario implies the existence of complex RNA cells that were able to produce sophisticated enzymes such as RNA polymerases, reverse transcriptases and ribonucleotide reductases during the late second age of the RNA world. This has been sometimes disputed, because the low fidelity of RNA replication would have limited the size of RNA genomes [13]. However, the fidelity of transcription has been probably underestimated. Recent work has shown that some RNA damage can be repaired [14,15] and that RNA polymerases fidelity is increased by specific transcription factors [16,17]. It is therefore difficult to extrapolate a maximum size limit for the genomes of ancient RNA cells from the maximum size limit of the genomes of modern RNA viruses. The latter are probably not representative of the genomes of ancient RNA cells, since viral RNA genomes have been most likely streamlined during their long period of co-evolution, first with RNA cells and later on with DNA cells. Furthermore, as viruses can find an advantage in low replication fidelity, the latter could well be a secondary feature of their replication apparatus. In agreement with this view, Pugachev et al. [18] were recently able to obtain high fidelity replication of viral RNA by in vivo selection. One can thus safely suggest that the RNA world of the late second age was characterised by a great diversity of cells with different lifestyle strategies and a great variability in genome sizes, gene numbers and chromosome architecture.

7. The U-DNA world as an intermediate step in the RNA/DNA transition

In modern cells, dTMP is produced by thymidylate synthases from dUMP and not by reduction of TTP (Fig. 3A). This strongly suggests that U-DNA (DNA with uracil instead of thymidine) was an intermediate in the RNA/DNA transition. Several authors have thus proposed that a U-DNA world preceded the present T-DNA world [13,19]. It is remarkable that some bacterial viruses have U-DNA genomes [20]. Such viruses could be relics of this U-DNA word in the same way as RNA viruses could be relics of the RNA world. Degradation of U-DNA in the U-DNA world and/or cytosine deamination would have produced the initial dUMP substrates for the first thymydilate synthase [21].

It is usually considered that invention of DNA was such an important event that it should have happened only once. However, it has been shown recently that two different families of structurally unrelated (non-homologous) thymidylate synthases are present in modern cells, ThyA and ThyX [22]. This indicates that thymidylate synthases were invented at

least twice independently in the U-DNA world. The fact that a crucial enzymatic activity of the RNA/DNA transition was invented at least twice indicates that a strong selection pressure drove this transition. Hence the critical question for all hypotheses dealing with the origin of DNA is: why was DNA invented after all?

8. Why was DNA invented and who did it? The viral hypothesis

It is usually considered that RNA was replaced by DNA in the course of evolution for two main reasons:

1°) DNA is more stable than RNA because the 2'O of the ribose is a very reactive atom that can attack the phosphodiester bond;

2°) deamination of cytosine into uracil (a common spontaneous chemical reaction) can be repaired in DNA but not in RNA (for obvious reasons!).

As a consequence, the substitution of RNA by DNA as cellular genetic material appears to be justified in order to allow genome size to increase in the course of evolution (the larger apparently the better). DNA-cells with enlarging genomes will become more complex and finally out-compete their RNA-based ancestors. However, this kind of argumentation is not valid from an evolutionary point of view. It's like to argue that feathers were invented in Dinosaurs in order to prepare for the future flight of birds! In a Darwinian scheme, one has to identify the selection pressure that first triggered the evolutionary process by bringing an immediate benefit to the organism in which the innovative mutation(s) appeared. An obvious selection pressure for an organism to modify its genome can be to protect it from attacks by hostile competitors. Chemical modification of its RNA genome into something "new", immune to RNAses, could have given an immediate benefit to an organism fighting for its life in the jungle of the second age of the RNA world. The principle of continuity suggests that such organism was an RNA virus, since we know that some modern DNA viruses have modified their genome precisely for this purpose (for instence via methylation, hydroxymethylation or even more complex chemical modification, see Ref. [23] for review). It is thus reasonable to think that viruses were also the promoters of previous genome modifications, from RNA to U-DNA, and from U-DNA to T-DNA [19,21].

Transformation of RNA into U-DNA would have made the genome of the first U-DNA virus resistant to all mechanisms invented by RNA cells to destroy the RNA of hostile viruses (in particular double-stranded DNA). Viruses that have conserved an RNA genome, have indeed evolved alternative mechanisms to protect their genetic material against RNA-degrading or modifying enzymes. Some of them keep their RNA genome into their capsid all along the infection process, whereas others encode proteins that inhibit cellular RNA degradation or modification mechanisms (e.g. demethylases). Considering the variety of strategies used by viruses to escape the host defence, it would be very surprising if genome modification (a very simple one) had not been used.

The idea that both ribonucleotide reductases and thymidylate synthases first appeared in viral genomes and were later on transferred to cells (see below) is compatible with phylogenetic analyses of these enzymes families that include a mixture of viral and cellular sequences. Many viruses encode ribonucleotide reductases or thymidylate synthases that are only distantly related to those encoded by their hosts [21]. In addition, although the direction of ancient transfers between cells and viruses are difficult to polarise, some recent transfers from viruses to cells can be clearly documented [21].

9. From simple to complex DNA viruses

The formation of DNA genomes probably occurred in several discrete steps leading to more and more complex structures and replication mechanisms [21]. In a first step, the simplest hypothesis is to imagine a single-strand RNA virus with a double-strand RNA replication intermediate becoming a single-stranded DNA virus with a DNA/RNA double-strand replication intermediate (Fig. 3B). In that case, the same enzyme could have originally retro-transcribed RNA into DNA and, in a second round, transcribed DNA into RNA. From this starting point, one can imagine that diversification of the initial DNA viral lineage give birth to DNA viruses replicating their DNA without RNA intermediate via the invention of DNA-dependent DNA polymerases, opening the possibility to produce double-strand DNA. The first double-strand DNA genomes were probably replicated by an asymmetrical mechanism (one strand after the other), like the genomes of double-strand RNA viruses and some DNA viruses. This mechanism only requires at first a DNA polymerase with strand-displacement activity. It can be made more efficient in a second step by a helicase and processivity factors to help the polymerase. Larger genomes could be then replicated more rapidly if the replication of the two strands became coupled. This requires a mechanism to produce primers to initiate replication of the lagging strand before completion of the leading strand, hence the invention of DNA primase. Finally, the rapid replication of very large genomes would become possible by coupling the replication of the lagging and leading strands. This would require the tight coordination of the helicase, primase and DNA polymerase into a replication factory.

10. The transfer of DNA from viruses to cells

Interestingly, one can find examples of all the different replication mechanisms discussed above in the present viral world. In contrast, all types of cells only use the more complex symmetric mechanism for DNA replication (Fig. 3B). To explain this, I would like to propose here that evolution of DNA replication mechanisms, from the simplest to the more

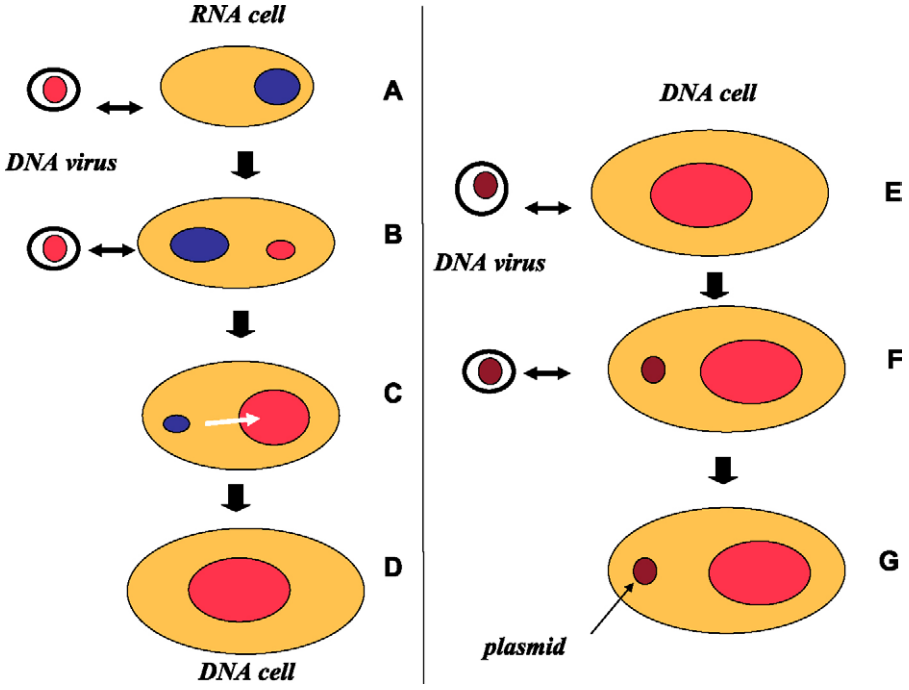

Fig. 4. A model for the transfer of DNA from viruses to cells for the origin of cellular DNA chromosomes and plasmids.
A DNA virus (DNA genome in red) infected an RNA-cell (RNA genome in blue) (A) and co-evolved with it in a carrier state (B). Genes from the cellular RNA genomes are progressively transferred to the viral DNA genome by retrotranscription (white arrow) and the viral genome evolved into a DNA plasmid of the RNA-cell (C). The DNA plasmid finally out-competed the RNA genome and become a cellular DNA chromosome (D). Infection of a DNA cell by a DNA virus can led, by a similar mechanism, to a DNA cell with both a plasmid and a chromosome (E–G). This scenario should produce a procaryotic type of cell. For the formation of eukaryotic cells, the nucleus could have originated by viral-induced recruitment of intracellular membranes to produce the nuclear membrane, by a mechanism derived from the process used by large double-stranded DNA viruses to form their envelopes.

complex one, occurred entirely in the viral world, and that transfer of DNA from viruses to modern cells only occurred at the end of the process. In other words, I suggest that both DNA and a complex DNA replication machinery (performing symmetric replication) were transferred at the same time from one or several (see below) large DNA viruses to cells. How can this have happened? Let's speculate a little bit more (Fig. 4). It is clear now that the majority of DNA viruses are not lytic or lysogenic, but live most of the time in a carrier state in their cellular hosts. This means that many ancient RNA cells of the second age should have contained, at least transiently, viral DNA genomes beside their own cellular RNA genomes. If the virus or the cell also encoded a reverse transcriptase, the DNA virus will be able to progressively integrate into its DNA genome cellular genes (formerly RNA genes) that could be helpful to facilitate the infection process and/or its survival into the cell (a rampant process still at work in modern viruses). At some stage, a virus might have been unable to accommodate anymore its enlarging genomes into a capsid; and the viral genomes finally became a DNA plasmid after loosing genes encoding capsid proteins or proteins involved in the infectious process. The replication of the new intracellular DNA genome of viral origin being more efficient than that of the ancient cellular RNA genome, the former could have finally taken over the complete cellular machinery by capturing all information needed for cellular life. The end of this process would be the complete elimination of the ancient cellular RNA genome and the formation of a modern cellular DNA chromosome (Fig. 4C).

The new cell with a DNA genome will rapidly diversify into large populations of DNA-cells that will easily outcompete populations of RNA cells for the two reasons usually mentioned to explain the predominance of DNA over RNA: much higher chemical stability and the possibility to design specific repair mechanism to counteract cytosine deamination. The possibility for DNA-cells to manage large genomes would have definitely give them a critical advantage in their competition with populations of RNA cells, leading to the complete elimination of all RNA-cell lineages.

Interestingly, the same scenario can explain the origin of plasmids from viruses if the host RNA-cell is replaced as starting point by an early DNA cell (Fig. 4D, E). Plasmids and viruses encode many homologous proteins, suggesting indeed a strong evolutionary relationship [21]. It is likely that plasmids evolved from viruses rather than the opposite since it was originally easier to loose genes encoding for capsid proteins and other structural elements for a virus living into a carrier state than to acquire them de novo for a plasmid. As DNA viruses were probably already diversified at the time of plasmid formation and transfer of DNA from viruses to cells (the latter leading to chromosome formation) the scenario proposed here would explain why plasmidic and viral replication origins are so diverse and strikingly different from cellular replication origins. On the other hand, the intricate

modular and dynamic evolution observed in modern populations of plasmids and viruses could be a relic of this early period of co-evolution between viral genomes, plasmids and cellular chromosomes.

11. The viral theory for the origin of DNA can explain the stunning diversity of DNA-associated proteins and their atypical phylogenetic distribution

In the viral theory for the origin of DNA, DNA and its associated proteins (here called DNA proteins) used to build up more and more sophisticated DNA replication mechanisms depicted in Fig. 3B were recruited by early emerging lineages of DNA viruses from proteins previously dealing with RNA [19,21]. These recruitments probably did not occur at once, but at different stages of DNA viral evolution and diversification. Accordingly, the same types of activities might have originated at different times independently in various viral lineages. For instance, several families of DNA viruses with simple asymmetric DNA replication, evolving toward semi-symmetric types of DNA replication, might have recruited independently different RNA helicases and RNA polymerases to engineer DNA helicases and DNA primases. If the transfer of DNA from viruses to cells indeed occurred at a late stage in the evolution of DNA replication mechanisms (as assumed previously) these independent recruitments would have occurred before the establishment of DNA cell lineages. This scenario allows to make several predictions that are indeed supported by data from genomics and protein structural analyses.

- 1°) Many DNA proteins performing analogous functions should not be homologous, since they originated independently in the first age of the DNA (viral) world. This is indeed the case of, among others, DNA polymerases, DNA primases, DNA helicases and DNA topoisomerases. In the extreme case of DNA polymerases, seven families have already been described that share no sequence similarities [24,25]. In a few cases (e.g. some DNA polymerase families), it is difficult to know if the absence of similarities really testify fort the absence of homology or if it is simply due to the loss of phylogenetic signal for very ancient divergence. However, in other cases (DNA primase, helicase or topoisomerases) the absence of homology is clearly established when enzymes with similar function exhibit no overall structural similarities and belong to different protein families (i.e. when they are phylogenetically more related to proteins with different functions). For an exhaustive study of this problem in the case of DNA replication proteins, see Ref. [26].
- 2°) Some families of homologous DNA proteins might have started diversifying in the viral world before being transferred independently to different cellular lineage. In that case, some cellular and viral proteins, although homologous, should exhibit no specific relationships between host and viruses. This prediction turned out to be valid with a few exceptions. In fact, in phylogenetic trees, most viral ribonucleotide reductases, thymidylate synthases, RNA or DNA polymerases, DNA ligases, DNA primase and so on, form various monophyletic groups that are only distantly related to their cellular counterparts [21,24,27–29].
- 3°) only a subset of all DNA proteins made during early DNA virus evolution would have been transferred to cells, meaning that most viruses should encode DNA proteins that have no cellular homologues. This is indeed the case for many DNA proteins encoded by viruses such as the T3/T7 type of RNA polymerases, the Herpes DNA primase, Rep proteins used to initiate proteins for rolling-circle replication, or else priming proteins for plasmid and viral DNA replication (Ref. [21,27]). A striking example is the recently discovered DNA polymerase (PolE) that is exclusively encoded by archaeal and bacterial plasmids [25]. The abundance of orphan genes in viral and plasmid genomes suggests that many novel DNA proteins remain to be discovered by systematically exploring the world of extra-chromosomal elements.

Up to now, I have only emphasised ancient transfers of DNA proteins from viruses to cells, but it is likely that such transfers have occurred continuously during evolution, and still occur at present. If viral DNA proteins are indeed so diverse, one should expect to find in some cellular lineages DNA proteins missing from all other ones (the product of a recent transfer of a specific viral protein into this particular lineage). This could be the case, for instance, for the DNA topoisomerase V of *Methanopyrus kandleri*. This atypical type I DNA topoisomerase has presently no detectable homologue in other Archaea nor in the two others domains ([30], and recent personal observation). DNA topoisomerase V corresponds to the association of a module distantly related to some integrases encoded by viruses or plasmids and a module that exhibits DNA repair activity in vitro. It is difficult to imagine that this complex protein was created de novo in the *M. kandleri* lineage. It is easier to think that it originated from a viral (plasmid) protein that was transferred only recently in the lineage leading to *M. kandleri*.

Many orphan proteins found in completely sequenced genomes thus could be of viral origin. The antiquity of viruses and their high rate of evolution would have produce an infinite reservoir of viral (and plasmidic) genes ready to be tested for use by cells, and to be integrated into cellular genomes if they confer some selective advantage.

12. The puzzling story of DNA replication diversity can be explained by the viral hypothesis

There are presently two cellular versions of the symmetric DNA replication mechanism, one in Bacteria, another one in Archaea/Eukarya [26,31,32]. Indeed, central components of these mechanisms (the replicative DNA polymerase, the DNA primase and the replicative helicase) are non-homologous.

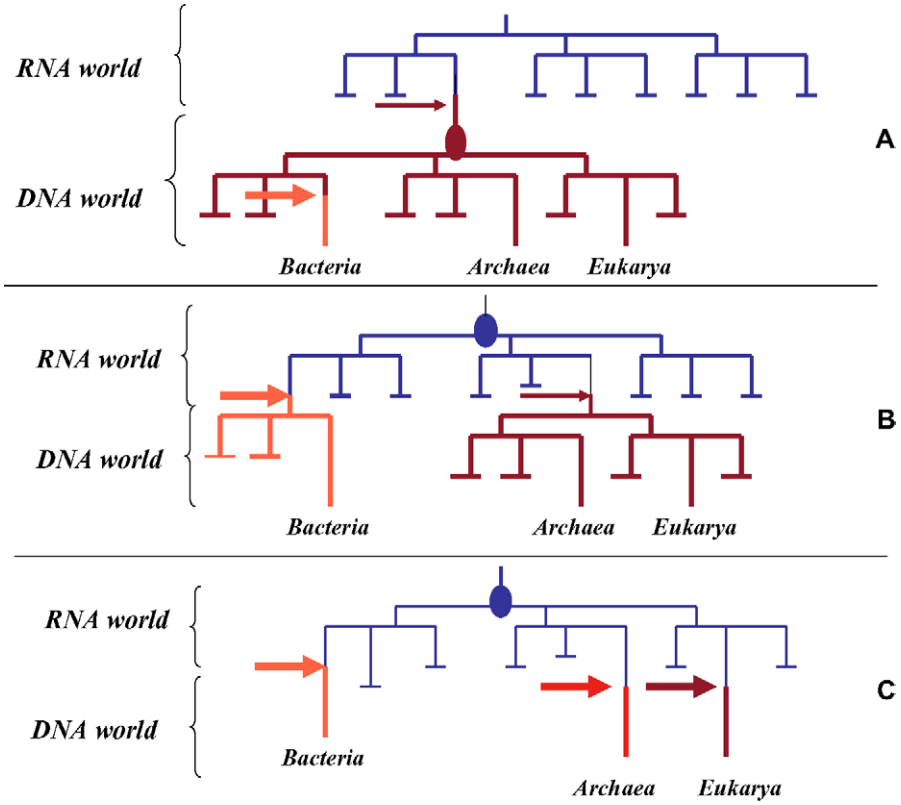

Fig. 5. Different scenarios for the transfer of two or three DNA replication mechanisms from viruses to cells.
RNA lineages are in blue and DNA lineages in red/orange/or brown. Open circle, LUCA. Red arrows: transfer of DNA replication mechanisms from viral to cellular lineages. A, a first mechanism (brown) was transferred before LUCA (DNA-LUCA) and was replaced later on by a second one (orange) [35], here in the bacterial branch. Another possibility, not illustrated here, is that such replacement occurred in a common lineage to Archaea and Eukarya. B: Two independent transfers from two different viruses after LUCA. C: Three independent transfers from different viruses after LUCA, each of them being at the origin of one cellular domain.

This has been originally explained by positing two independent inventions of symmetric DNA replication mechanisms, either before or after LUCA [32,33] (Fig. 5). In the first case, both were present in LUCA and later on differentially lost in the three different domains, whereas in the second case, the invention of DNA replication occurred once in the bacterial lineage and another time in a lineage common to Archaea and Eukarya. Alternatively, the ancestral mechanism present in LUCA could have been displaced by another one, either in Bacteria or in a lineage common to Archaea and Eukarya (non-orthologous displacement, [34]).

In these hypotheses, it was not clear why DNA replication originated twice independently in the lineage leading to LUCA and what was the origin of the mechanism that displaced the ancestral one in one or two domains. The viral hypothesis for the origin of DNA readily gives an answer to these two questions. The existence of several non-homologous DNA replication mechanisms now makes sense, since DNA replication mechanisms should have emerged independently in different viral lineages. Furthermore, these different viral mechanisms became obvious source of possible material for non-orthologous displacement [19,35,36].

Originally, the viral hypothesis for the origin of cellular DNA replication mechanisms postulated either one transfer of DNA from viruses to cells before LUCA (the DNA-LUCA version, Fig. 5A) or two independent transfers after LUCA (the RNA-LUCA version). In the first scenario, (DNA-LUCA, Fig. 5B) the transfer of a second DNA replication mechanism (with elimination of the first one) should have occurred after LUCA, either in the bacterial lineage (Fig. 5B) or in a lineage common to Archaea and Eukarya. In the second scenario (RNA-LUCA), one mechanism was transferred to the bacterial lineage and the second one in the lineage common to Archaea and Eukarya. Here, I would like to explore an alternative hypothesis, i.e. the possibility that the transfer of different DNA replication mechanisms to RNA cells of the second age coincided with the origin of the three cellular domains that we know today (Figs. 5C and 6).

13. The three viruses–three domains hypothesis

It has been noticed for a long time by Carl Woese that the tempo of evolution was higher at the time of the universal ancestor (LUCA) than it is now [37,38]. This can be inferred from the observation that most homologous informational proteins present in the three domains (e.g. ribosomal proteins) are strikingly different in term of sequence from one domain to another (each exhibiting a canonical pattern *sensu* Woese), while they are very similar across all phyla of a given domain

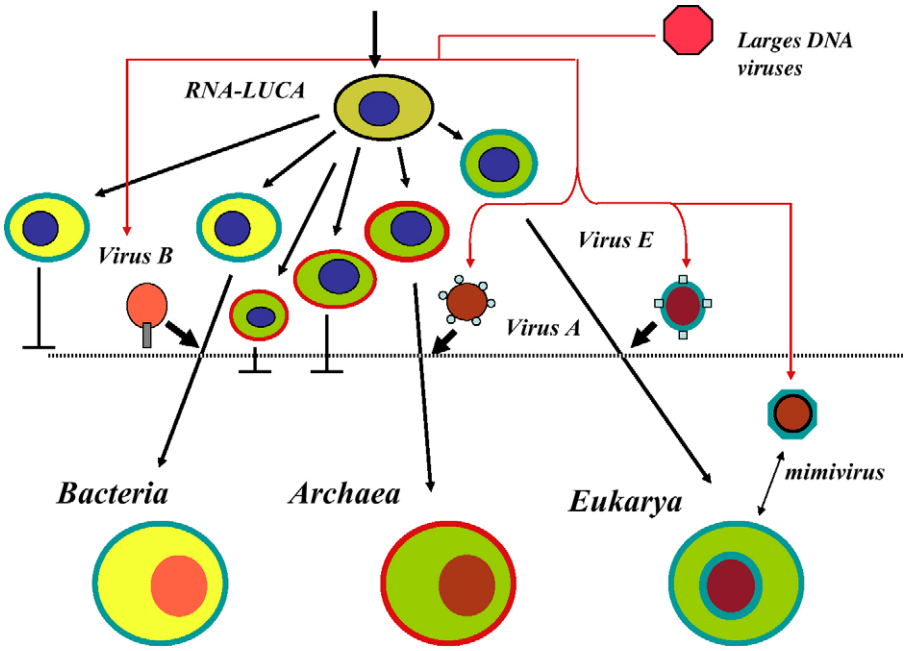

Fig. 6. Formation of the three domains from three independent DNA transfers from viruses to RNA cells.
Many lineages of RNA cells (blue genomes) with various types of membranes (red or blue for archaeal-like and bacterial-like lipids, respectively) and translation apparatus with different canonical patterns (from yellow to green) diverged from an RNA-based LUCA during the second age of the RNA world. In three of them, the RNA genomes were replaced independently by DNA genomes from different large double-stranded DNA viruses (A–C) (see Fig. 7). The DNA viruses and the RNA cells at the origin of Archaea and Eukarya shared more similar features in their informational apparatus (symbolised by more similar colours), but the RNA cells at the origin of Bacteria and Eukarya shared similar lipids (blue). The descendents of the three ancestors of the present-day domains eliminated all RNA cells lineages. The dashed line indicates the time of the transition from RNA-to-DNA genomes, leading to the drastic reduction in evolutionary tempo [37].

(although intra-domain diversification occurred probably during a much longer period of time than the original diversification of the three domains from LUCA). This indicates that the evolutionary rates of these proteins dramatically decreased shortly after each domain originated and started to diversify into different phyla. One should thus postulate that a critical event was responsible for the transition between these two periods of evolution. In the framework of the viral scenarios suggested here, an intriguing possibility is that this critical event was simply the transfer of DNA from viruses to cells. In that "three viruses–three domains" hypothesis, Archaea, Bacteria and Eukarya originated from different RNA cells lineages which acquired independently their DNA genomes from three different DNA viruses; as a consequence, LUCA was an RNA-cell of the second age, in agreement with earlier suggestion by Woese [39] who call it a progenote (Fig. 6).

The proposal that each cellular domain originated from a different DNA virus might be counter-intuitive for most readers (in a way even for myself). However, this "three viruses–three domains" hypothesis explains some puzzling observations that are not readily take into account by other scenarios. For instance, it is not clear from traditional hypotheses why Archaea and Eukarya share homologous informational proteins, whilst Bacteria and Eukarya share homologous membrane structure and composition. Similarly, in hypotheses suggesting that Eukaryotes originated from the merging of an archaeon (the proto-nucleus) with a bacterium (the proto-cytoplasm), such as in the recent ring of life paper [40], it is not clear how the original bacterial/archaeal double-membrane of the nucleus evolved into a single membrane folded onto itself (the present-day nuclear membrane). In the framework of the "three viruses–three domains" hypothesis, since a great diversity of RNA-cell lineages should have existed at that time, one can simply imagine that the RNA-cell ancestor of Eukaryotes had "bacterial-like" lipids but "archaeal-like" ribosomes (Fig. 6).

All traditional scenarios for the origin of the three domains also failed to explain why the eukaryal and archaeal DNA replication machineries, although similar, exhibit critical differences. In particular, archaeal and eukaryal DNA polymerases are not specifically related, but both originated from different viral groups in molecular phylogenies [24]. Similarly, major eukaryal DNA topoisomerases are unrelated to archaeal ones [28,41]. It is therefore difficult to explain how the eukaryotic system evolved from the archaeal one (the classical view). In contrast, the differences observed between the archaeal and eukaryal replication machineries make sense in the framework of the "three viruses–three domains" hypothesis since Archaea and Eukarya received different complements of DNA polymerases and topoisomerases from different viruses.

In the case of Eukarya, the hypothesis proposed here is reminiscent of the recent suggestion by Takemura [42] and Bell [43] that the eukaryotic nucleus originated from a large DNA double-stranded virus. In their original proposal, these authors suggested that eukaryotes evolved from an archaeon infected by a virus related to present-day Poxviruses. How-

ever, this cannot be easily reconciled with the complete absence of archaeal-type membranes in present-day eukaryotes. A more plausible version of their idea is that the putative viral ancestor of the nucleus was not present in an archaeon but in a cell (either RNA or DNA) with eucaryotic-type membranes. In the scenario suggested here, the nature of the infecting virus and of its interaction with the host RNA-cell could have determined the type of cellular organisation of the newly emerging DNA cell, either prokaryote (in Archaea and Bacteria) or eukaryotes. Briefly, transfer of DNA from a "simple" virus could have led to a prokaryotic cell, whereas transfer of DNA from a complex virus could have led to a eukaryotic cell, if the virus used to recruit intracellular membranes of the RNA-cell for the formation of its envelope. However, further elaboration of such models for the origin of the eukaryotic nucleus will have to consider that nucleated bacteria also exist [44]. Deciphering the relationships between eukaryotic and bacterial nuclei will be essential to understand the logic of the prokaryote/eukaryote transition(s) (for comments on a recent meeting on this topic, see Ref. [45]).

Interestingly, the "three viruses–three domains" hypothesis can nicely explain the phylogeny of a crucial enzyme for the RNA-to-DNA world transition, the DNA-dependent RNA polymerase. Homologous RNA polymerases are present both in the three cellular domains and in a large group of double-stranded DNA viruses that include, among others, Poxviruses and the recently described giant Mimivirus [29]. Surprisingly (for traditional, hypotheses), the cellular RNA polymerases are interspersed with RNA polymerases of these "eucaryotic" viruses in a phylogenetic tree of bacterial RNA polymerase β subunits homologues (Fig. 7). The Mimivirus RNA polymerase branches in-between Archaea and Eukaryotes, whilst the RNA polymerases of other large DNA viruses branch in-between Bacteria and a clade comprising Archaea, Eukarya and the Mimivirus (Fig. 7). This phylogeny is strikingly compatible with the "three viruses–three domains hypothesis", suggesting that the three DNA viruses at the origin of the three cellular domains were large DNA viruses encoding homologous RNA polymerases (Fig. 6). Of course, one can also explain this result by the transfer of RNA polymerases from now extinct cellular lineages to ancestors of present-day large DNA viruses. However, the idea that the "first specific common ancestor" of each domain originated from the transition of an RNA-cell to a DNA cell (via a virus) has a greater power of explanation. First, as previously mentioned, it immediately explains why the evolutionary tempo was drastically reduced at the onset of the formation of each domain, since DNA genomes can indeed be replicated more faithfully than RNA genomes. Furthermore, since all RNA-cell lineages became probably rapidly extinct after these three events, the hypothesis explains why there are only three cellular domains with clear canonical patterns and no intermediate situation. In this scenario, evolution of life from the second age of the RNA world to the DNA world has erased the possibility of further similar transitions, preventing the formation of new cellular domains.

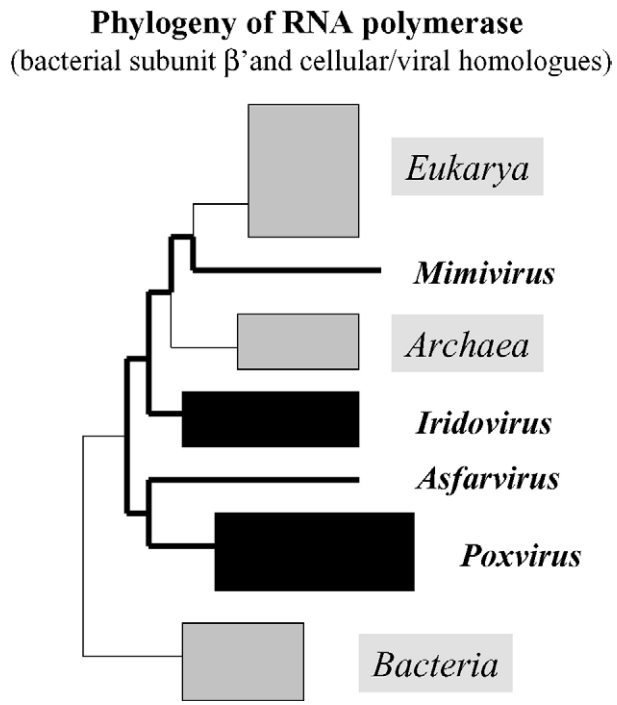

Fig. 7. Schematic phylogeny of homologous RNA polymerases largest subunit encoded by cellular and double-stranded large DNA viruses.
Cellular domains are in grey while viral groups and branching are in black. The topology and branch lengths are adapted from a neighbour-joining tree published by Raoult and coworkers ([38], supplementary material). This tree encompasses homologues of the bacterial β′ subunit of the RNA polymerase.

14. Conclusion

RNA viruses have been considered for a long time as possible relics of the RNA world. However, they were considered (with few exceptions) as passive witnesses of this period (fragments of RNA cells that escaped to become parasites) and DNA viruses were never considered as possible players in a world of RNA cells. Comparative genomic now forces us to propose new scenarios to explain the diversity and puzzling phylogenetic distribution of proteins involved in the transition from the RNA to the DNA world and the very origin of DNA genomes. In the scenarios suggested here, both RNA and DNA viruses were very active forces in early cellular evolution. In this review I have only considered in some detail the late RNA world, however, it is possible that dynamic interactions between RNA viruses and RNA cells also played an important role in early steps of RNA-cell evolution. Exploration of the dynamic interactions between RNA viruses and the present RNA world with an evolutionary oriented mind could be one way to put the above ideas under experimental testing.

Acknowledgments

I thank Marie-Christine Maurel to invite me participating to the 5th sifrARN meeting in Arcachon and Simonetta Grib-

aldo for discussion and correction of this manuscript. Work on DNA replication in my laboratory are supported by Grants and Fellowships from the Association de Recherche contre le Cancer (ARC) and the Human Frontier Science Program (HFSP).

References

[1] L. Ribas de Pouplana, The Genetic Code and the Origin of Life, Landes Bioscience, 2004, pp. 1–253.
[2] W. Gilbert, The RNA world, Nature 319 (1986) 618.
[3] D.C. Jeffares, A.M. Poole, D. Penny, Relics from the RNA world, J. Mol. Biol. 46 (1998) 18–36.
[4] Kandler, The early diversification of life and the origin of the three domains, a proposal, in: J. Wiegel, M.W.W. Adams (Eds.), Thermophiles: The Keys to Molecular Evolution and the Origin of Life, Taylor and Francis, 1998, pp. 19–28.
[5] Y. Koga, T. Kyuragi, M. Nishihara, N. Sone, Archaeal and bacterial cells arise independently from noncellular precursors? A hypothesis stating that the advent of membrane phospholipid with enantiomericglycerophosphate backbones caused the separation of the two lines of descent, J. Mol. Evol. 46 (1998) 54–63.
[6] W. Martin, M.J. Russell, On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells, Philos. Trans. R. Soc. London B Biol. Sci. 358 (2003) 59–83.
[7] J. Pereto, P. Lopez-Garcia, D. Moreira, Ancestral lipid biosynthesis and early membrane evolution, Trends Biochem. Sci. 29 (2004) 469–477.
[8] P.B. Moore, T.A. Steitz, The involvement of RNA in ribosome function, Nature 418 (2002) 229–235.
[9] C.R. Woese, The Genetic Code: The Molecular Basis of Genetic Expression, Harper and Row, New York, 1967.
[10] P. Forterre, New hypotheses about the origins of viruses, prokaryotes and eukaryotes, in: J.K. Trân Thanh Vân, J.C. Mounolou, J. Shneider, C. Mc Kay (Eds.), Frontiers of Life, Editions Frontières, Gif-sur-Yvette, France, 1992, pp. 221–234.
[11] M. Kolberg, K.R. Strand, P. Graff, K. Andersson, Structure, function and mechanism of ribonucleotide reductases, Biochim. Biophys. Acta 1699 (2004) 1–34.
[12] S.J. Freeland, R.D. Knight, L.F. Landweber, Do proteins predate DNA? Science 286 (1999) 690–692.
[13] A. Poole, D. Penny, B.-M. Sjöberg, Methyl-RNA: an evolutionary bridge between RNA and DNA? Chem. Biol. 7 (2000) 207–216.
[14] P.A. Aas, M. Otterlei, P.O. Falnes, C.B. Vagbo, F. Skorpen, M. Akbari, O. Sundheim, M. Bjoras, G. Slupphaug, E. Seeberg, H.E. Krokan, Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA, Nature 421 (2003) 859–863.
[15] P.O. Falnes, Repair of 3-methylthymine and 1-methylguanine lesions by bacterial and human AlkB proteins, Nucleic Acids Res. 32 (2004) 6260–6267.
[16] H. Koyama, T. Ito, T. Nakanishi, N. Kawamura, K. Sekimizu, Transcription elongation factor S-II maintains transcriptional fidelity and confers oxidative stress resistance, Genes Cells 8 (2003) 779–788.
[17] U. Lange, W. Hausner, Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage, Mol. Microbiol. 52 (2004) 1133–1143.
[18] K.V. Pugachev, F. Guirakhoo, S.W. Ocran, F. Mitchell, M. Parsons, C. Penal, S. Girakhoo, S.O. Pougatcheva, J. Arroyo, D.W. Trent, Monath TP high fidelity of yellow fever virus RNA polymerase, J. Virol. 78 (2004) 1032–1038.
[19] P. Forterre, The origin of DNA genomes and DNA replication proteins, Curr. Opin. Microbiol. 5 (2002) 525–532.
[20] I. Takahashi, J. Marmur, 1: Replacement of thymidylic acid by deoxyuridylic acid in the deoxyribonucleic acid of a transducing phage for *Bacillus subtilis*, Nature 197 (1963) 794–795.
[21] P. Forterre, J. Filee, H. Myllykallio, Origin and evolution of DNA and DNA replication machineries, in: L. Ribas de pouplana (Ed.), The Genetic Code and the Origin of Life, Landes Bioscience, 2004 pp. 145–168.
[22] H. Myllykallio, G. Lipowski, D. Leduc, J. Filee, P. Forterre, U. Liebl, An alternative flavin-dependent mechanism for thymidylate synthesis, Science 297 (2002) 105–107.
[23] R.A.J. Warren, Modified bases in bacteriophage DNAs, Annu. Rev. Microbiol. 34 (1980) 137–158.
[24] J. Filee, 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.
[25] G. Lipps, S. Rother, C. Hart, G. Krauss, A novel type of replicative enzyme harbouring ATPase, primase and DNA polymerase activity, EMBO J. 22 (2003) 2516–2525.
[26] D.D. Leipe, L. Aravind, E.V. Koonin, Did DNA replication evolve twice independently? Nucleic Acids Res. 27 (1999) 3389–3401.
[27] J. Filée, P. Forterre, J. Laurent, The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies, Res. Microbiol. 154 (2003) 237–243.
[28] D. Gadelle, J. Filee, C. Buhler, P. Forterre, Phylogenomics of type II DNA topoisomerases, Bioessays 25 (2003) 232–242.
[29] D. Raoult, S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. La Scola, M. Suzan, J.M. Claverie, The 1.2-megabase genome sequence of Mimivirus, Science 306 (2004) 1344–1350.
[30] G.I. Belova, R. Prasad, I.V. Nazimov, S.H. Wilson, A.I. Slesarev, The domain organization and properties of individual domains of DNA topoisomerase V, a type 1B topoisomerase with DNA repair activities, J. Biol. Chem. 277 (2002) 4959–4965.
[31] G.J. Olsen, C.R. Woese, Archaeal genomics: an overview, Cell 89 (1997) 991–994.
[32] A.R. Mushegian, E.V. Koonin, A minimal gene set for cellular life derived by comparison of complete bacterial genomes, Proc. Natl. Acad. Sci. USA 93 (1996) 10268–10273.
[33] D.R. Edgell, W.F., Doolittle Archaea and the origin(s) of DNA replication proteins, Cell 89 (1997) 995–998.
[34] E.V. Koonin, A.R. Mushegian, P. Bork, Non-orthologous gene displacement, Trends Genet. 12 (1996) 334–336.
[35] 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.
[36] L.P. Villarreal, A. DeFilippis, Hypothesis for DNA viruses as the origin of eukaryotic replication proteins, J. Virol. 74 (2000) 7079–7084.
[37] C.R. Woese, Bacterial evolution, Annu. Rev. Microbiol. 270 (1987) 221–271.
[38] C.R. Woese, Interpreting the universal phylogenetic tree, Proc. Natl. Acad. Sci. USA 97 (2000) 8392–8396.
[39] C.R. Woese, The primary lines of descent and the universal ancestor, in: D.S. Bendall (Ed.), Evolution from Molecules to Men, Cambridge Universoty Press, Cambridge, 1983, pp. 209–233.
[40] M.C. Rivera, J.A. Lake, The ring of life provides evidence for a genome fusion origin of eukaryotes, Nature 431 (2004) 152–155.
[41] M.C. Serre, M. Duguet, Enzymes that cleave and religate DNA at high temperature: the same story with different actors, Prog. Nucleic Acid Res. Mol. Biol. 74 (2003) 37–81.
[42] M. Takemura, Poxviruses and the origin of the eukaryotic nucleus, J. Mol. Evol. 52 (2001) 419–425.
[43] P.J.L. Bell, Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? J. Mol. Evol. 53 (2001) 251–256.
[44] M.R. Lindsay, R.I. Webb, M. Strous, M.S. Jetten, M.K. Butler, R.J. Forde, J.A. Fuerst, Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell, Arch. Microbiol. 175 (2001) 413–429.
[45] E. Pennisi, The birth of the nucleus, Science 305 (2004) 766–768.