MicroOpinion

Displacement of cellular proteins by functional analogues from plasmids or viruses could explain puzzling phylogenies of many DNA informational proteins

Patrick Forterre

Institut de Génétique et Microbiologie, Bat 409, CNRS, UMR 8621, Université Paris-Sud, 91405 Orsay Cedex, France.

Summary

Comparative genomics has revealed many examples in which the same function is performed by unrelated or distantly related proteins in different cellular lineages. In some cases, this has been explained by the replacement of the original gene by a paralogue or non-homologue, a phenomenon known as non-orthologous gene displacement. Such gene displacement probably occurred early on in the history of proteins involved in DNA replication, repair, recombination and transcription (DNA informational proteins), i.e. just after the divergence of archaea, bacteria and eukarya from the last universal cellular ancestor (LUCA). This would explain why many DNA informational proteins are not orthologues between the three domains of life. However, in many cases, the origin of the displacing genes is obscure, as they do not even have detectable homologues in another domain. I suggest here that the original cellular DNA informational proteins have often been replaced by proteins of viral or plasmid origin. As viral and plasmid-encoded proteins are usually very divergent from their cellular counterparts, this would explain the puzzling phylogenies and distribution of many DNA informational proteins between the three domains of life.

Introduction

Most viral and plasmid-encoded proteins involved in central information-processing systems, such as polymerases, ligases and topoisomerases, are very different in terms of sequence similarity from their cellular functional analogues (for example, see Kotani et al., 1987; Bernstein and Bernstein, 1989; Forterre, 1992; Ilyina and Koonin, 1992; Kempken et al., 1992; Koonin, 1993; Pellock et al., 1996). In some cases, the viral or plasmid-encoded proteins have no detectable cellular homologues at all (e.g. bacteriophage RNA polymerases or proteins involved in rolling circle replication), whereas in other cases, they are only distantly related to their cellular homologues (e.g. DNA polymerases, DNA topoisomerases). This is usually explained by the rapid evolution of many (if not all) viral genes (Koonin, 1993). Another non-exclusive possibility is that some viral and plasmid genes encoding informational proteins diverged from their cellular counterparts very early on (Forterre, 1992; Hendrix et al., 1999), possibly even before the separation of the three domains of life (Forterre, 1992). This could explain why some viral and plasmid-encoded proteins are more related to viral, plasmid or cellular proteins from another domain than to cellular proteins from the domain of their host. For example, many T4 bacteriophage proteins are more similar to their eukaryal than to their Escherichia coli counterparts (Bernstein and Bernstein, 1989; Forterre, 1992), and DNA polymerases from some bacteriophages, such as the Bacillus subtilis phage Φ29, are more similar to those from some eukaryal viruses and plasmids than to B. subtilis DNA polymerases (Pellock et al., 1996). I realized recently that, whatever its cause (rapid evolution and/or early divergence), the great evolutionary distance between DNA informational proteins from viruses and plasmids and their cellular functional analogues could explain a puzzling aspect of microbial evolution, i.e. the huge evolutionary gap between DNA informational proteins from bacteria and their functional analogues in archaea and eukarya (Edgell and Doolittle, 1997; Olsen and Woese, 1997). My proposal is that many cellular DNA informational proteins have been displaced by viral or plasmid-encoded proteins shortly after the divergence of the three domains of life from LUCA. Before discussing this hypothesis in more detail, one should first consider to what extent non-orthologous gene displacements have
indeed been involved in the history of DNA informational proteins.

Non-orthologous gene displacement in the evolution of DNA informational proteins

Most DNA replication proteins that are functionally analogous in bacteria and eukarya have either no sequence similarities between these two domains or only limited similarities restricted to a few amino acids involved in active sites. For example, the polymerization subunit of the bacterial replicase, DNA polymerase III, has no sequence similarity at all to its eukaryotic counterparts, the catalytic subunits of DNA polymerases α, δ and ε. In fact, bacterial and eukaryotic replicases belong to two distinct DNA polymerase families, C and B, respectively (Ito and Braithwaite, 1991). Similarly, although the bacterial and eukaryotic initiator proteins (DnaA and ORC/CDC6 respectively (Ito and Braithwaite, 1991). To a lesser extent, a similar situation can be observed in repair/recombination and transcription systems, as some bacterial proteins involved in these processes have no homologues in eukarya, and vice versa (Aravind et al., 1999). In contrast, most functionally analogous proteins involved in translation (e.g. ribosomal proteins, elongation factors, tRNA synthetases) are clearly orthologues in bacteria and eukarya.

From comparative genomics (Brown and Doolittle, 1997; Olsen and Woese, 1997; Aravind et al., 1999), it turns out that all bacterial replicative proteins and many bacterial proteins involved in repair, recombination and transcription also have no detectable homologues in archaea. In contrast, some eukaryotic proteins involved in these mechanisms have readily detectable archaeal homologues, which are most probably their orthologues.

To explain why DNA replicative proteins were so different in bacteria on one side and in eukarya/archaea on the other, Koonin and colleagues suggested that LUCA was a member of the RNA world and that DNA informational proteins originated independently in the bacterial and eukaryal (archaeal) lineages from unrelated RNA informational proteins (Mushegian and Koonin, 1996; Aravind et al., 1999). However, not all DNA informational proteins are unrelated between bacteria and the two other domains. Some of them are indeed clearly homologues in all extant cellular organisms (e.g. large subunits of DNA-dependent RNA polymerases, ribonucleotide reductases, Topo IA, recombines of the RecA family and a few other DNA recombina-
some segregation, and its functional counterparts in eukaryotes, Topo II, belong to different DNA topoisomerase II families (Bergerat et al., 1997). Similarly, eukaryal Topo IB, which is involved in the relaxation of positive superturn at the eukaryotic replication fork, has no homologues in archaeal genomes. Such an erratic pattern of relationships between archaeal and eukaryal replicative proteins cannot be explained by the LUCA-RNA theory because, as many archaeal and eukaryal replicative proteins are most probably orthologues, the common ancestor of archaea and eukarya was certainly a DNA-based organism. It cannot be explained either by differences in evolutionary rates, as, for example, eukaryotic Topo IB (a type I DNA topoisomerase) is not even structurally and mechanistically related to its bacterial and archaeal proteins that perform the same function in DNA replication (which are both type II DNA topoisomerases) (Forterre et al., 1994). In that case, it is clear that the puzzling distribution of DNA topoisomerase between archaea and eukarya can only be explained by non-orthologous gene displacement.

The term non-orthologous gene displacement has been coined recently by Koonin to describe the presence of non-orthologous proteins (unrelated or paralogues) for the same function in different organisms (Mushegian and Koonin, 1996; Koonin and Galperin, 1997; Koonin et al., 1996; 1997) (Fig. 1A). Now that several completely sequenced archaeal and bacterial genomes are available, it is clear that the displacement of proteins responsible for essential functions by evolutionary unrelated or distantly related proteins has been extensive in the archaeal and eukaryal domains, and even more so between domains (Koonin et al., 1997; Doolittle, 1998a). A priori, one would have thought that non-orthologous gene displacement should be limited to proteins that do not physically interact with other proteins, because it is difficult to envisage the replacement of a protein that physically interacts with several partners by a phylogenetically distinctly related or unrelated protein (Jain et al., 1999). However, although DNA informational proteins are often part of macromolecular complexes, many well-documented cases of non-orthologous displacement between various lineages of a single domain have now been reported for genes encoding such proteins. For example, most of the primosome components are non-orthologous between E. coli and B. subtilis, whereas one of them (PriA) and other parts of the replication apparatus (DNA helicase, primase, replicase) are clearly orthologues (Bruand et al., 1995; Kunst et al., 1997). E. coli and B. subtilis also use analogous but non-orthologous systems to produce single-stranded DNA for genetic recombination, the RecBCD and the AddAB/RexAB helicase/exonuclease respectively (El Karoui et al., 1998), whereas they both use orthologous RecA to complete the recombination pathway. This example is highly significant as some Gram-positive bacteria possess the RecBCD system, whereas others have the AddAB/RexAB system, indicating that non-orthologous displacement of DNA informational proteins even occurred during the diversification of Gram-positive bacteria. The absence of homologues of the eukaryotic DNA replication initiator proteins ORC1/ CDC6 in the genome of the archaean Methanococcus jannaschii also suggests recent non-orthologous displacement. Indeed, this protein, which is present in all other completely sequenced archaeal genomes, probably plays an essential role in the initiation of archaeal DNA replication (Lopez et al., 1999a). Accordingly, its function should have been taken over by another unknown protein in M. jannaschii (Bernander, 1998). In all these cases, one cannot explain the puzzling pattern of protein distribution observed inside one domain by divergent rates of evolution or by the RNA-LUCA hypothesis! If non-orthologous gene displacement of DNA informational proteins thus clearly occurred between both archaea and eukarya (as in the case of type II DNA topoisomerases) and after the diversification of each domain in multiple lineages, I think

Fig. 1. A. Definition of orthologues and paralogues (Fitch, 1970): genes (or proteins) are orthologues in two species (A and B) if they descend from the same common ancestral gene in the last common ancestor of these two species. They are paralogues if they descend from homologous genes that were already duplicated in this common ancestor. In this example, black genes in A and B are orthologues (as well as white genes in A and B), whereas the black gene in A and the white gene in B are paralogues (and vice versa).

B. Non-orthologous gene displacement: one of the two orthologous genes can be displaced by a gene encoding a functionally analogous protein that can be either a paralogue (species A) or unrelated (non-homologue) to the displaced protein (species B, black square).
it most likely that non-orthologous gene displacement also occurred between bacteria and archaea/eukarya in the replicative, repair, recombination and transcription apparatus, explaining why some DNA informational proteins are unrelated between bacteria and archaea/eukarya. This hypothesis is more parsimonious than either the RNA-LUCA hypothesis or the replicative protein fast evolution hypothesis, as it gives the same explanation (non-orthologous gene displacement) for the presence of phylogenetically unrelated functional analogues between domains and between different lineages of a domain, whereas competitive hypotheses involved two unrelated explanations.

The idea that non-orthologous gene displacement in general has played a major role in the history of DNA informational proteins is supported by many examples of functional complementation that have been observed experimentally in studying DNA replication. For example, while the removal of RNA primers from Okazaki fragments is normally performed in *E. coli* by the 5′ to 3′ exonuclease activity of DNA polymerase I, it can be done by RNase H in mutants lacking this exonuclease activity (Ogawa and Okazaki, 1984). Similarly, *E. coli* DNA polymerase II can be used as replicase in some *E. coli* DNA polymerase III mutants (Rangarajan et al., 1997). The participation of DNA polymerase II at the *E. coli* replication fork is in line with the observation that, despite belonging to different DNA polymerase families (B and C respectively), both *E. coli* DNA polymerases II and III can interact in vitro with the bacterial processivity factor 8-clamp and clamp-loading factors (Bonner et al., 1992). Finally, a well-known example of functional complementation is the displacement of a thermosensitive *E. coli* DnaA protein by the initiator protein of another replicon at non-permissive temperature (Tresguerres et al., 1975). In that case, the *E. coli* mutant is rescued using the replication origin of an integrated replicon as a new oriC.

### Replicative viral and plasmid proteins as a bountiful source of displacing proteins

In some cases of non-orthologous gene displacement, the origin of the displacing gene is not a problem, as orthologues of this gene are known in other lineages or domains. For example, the displacement of class II lysyl-tRNA synthetases by those from non-homologous class I in some bacteria, or vice versa, can easily be explained by lateral gene transfer across domains, as both types are present in archaea and/or eukarya (Koonin and Aravind, 1998; Ibba et al., 1999). However, in many cases, the origin of the displacing gene remains obscure. This is especially striking for the DNA replication apparatus, because the displacing proteins often have no detectable homologue in the other domains. For example, as there is no member of the DNA polymerase C family in archaea and eukarya, then where did bacterial replicases come from? This problem probably explains why the hypothesis of non-orthologous gene displacement was not proposed earlier to explain the complex history of DNA informational proteins.

I suggest here that non-orthologous displacement of a cellular protein by a distantly related or unrelated viral or plasmid-encoded protein (Fig. 2) could explain those non-orthologous displacements whose origin is unclear.

In non-orthologous gene displacement, the genes encoding the displaced and the displacing proteins might have been present on the same chromosome for a long time, but the displacing gene might also have been captured from another cell by horizontal gene transfer. Considering the importance of viruses and plasmids as major agents of lateral gene transfer (Lawrence and Ochman, 1998), possibly even between phylogenetically remote lineages (Hendrix et al., 1999), they should play a significant role in such displacement. There is no reason to imagine that viruses and plasmids were only used in this process as vehicles for cellular genes. It is most likely that proteins encoded by viruses and plasmids were also often the displacing proteins themselves. Cells from all types of organisms contain a plethora of cryp776 viral and plasmid genes that have been integrated into chromosomal DNA (Lawrence and Ochman, 1998; Hendrix et al., 1999). Thus, there is plenty of opportunity for a cryptic viral or plasmid gene to displace a cellular gene that has been mutated, providing that the new protein can be successfully integrated into the network of pre-existing cellular proteins.

This idea has, in fact, already been suggested by Bouché et al. (1990), who discussed possible cases of co-option of bacteriophage-encoded genes by *E. coli*. An experimental example even exists, as it has been shown that *E. coli* recBC mutants can be rescued by mutations that activated the repair genes recE and recT encoded by the cryptic Rac prophage (Willis et al., 1985). Other bacterial repair proteins that might have prophage origin are RuvC and RusA (Bidnenko et al., 1998; Aravind et al., 1999). These proteins are only present in a subset of bacterial lineages, indicating recent invasion of the bacterial genome.

Interestingly, many bacterial replicative proteins that have no detectable homologues in the two other domains have viral or plasmid homologues (Table 1), supporting the idea that they might be of viral or plasmid origin. An especially striking case is that of bacterial replicase as, in addition to the genes dnaE and polC, the genome of *B. subtilis* contains a third gene encoding a parologue of the DNA polymerase III catalytic subunit, called yorL, which is located inside a cryptic prophage-like element. Bacterial DNA topoisomerase III are other putative candidates for a viral/plasmid origin, as they are only present in a limited number of bacteria (not shown) but have many homologues encoded by bacterial plasmids (Table 1).
Displacement of bacterial proteins involved in chromosome partitioning and segregation by viral or plasmid proteins could also explain major differences between these systems in different organisms (see discussion in Gérard et al., 1998). For example, the partition protein Spo0J, which plays an essential role in *B. subtilis* chromosome segregation, has no orthologue in *E. coli* (Lewis and Errington, 1997). A few archaeal and eukaryal DNA replication proteins also have viral or plasmid homologues, such as DNA polymerase of the B family and Topo IIA (Forterre, 1992). An interesting case is the eukaryal Topo IB, which, as mentioned previously, has no homologue in bacteria and archaea but, rather, one encoded by the Vaccinia virus.

The traditional view is that viral or plasmid-encoded proteins that have cellular homologues have been recruited from the cellular host. This is probably true in some cases, for example the DnaB homologue (protein Ban) of the *E. coli* bacteriophage P1 is much more closely related to *E. coli* bacteriophage P1 in phage-like element (Bacillus). But in other cases, there is no specific relationship between the viral protein and homologous proteins from their particular host. In fact, if one considers the likely hypothesis that exchange between viruses and cells can go both ways, and taking into account the great diversity of viral and plasmid proteins compared with cellular ones, the numerous viral (plasmid) genes available to the genomic cellular environment will have the tendency to displace the unique cellular genes originally inherited from LUCA. This would be similar to the ratchet mechanism proposed by Doolittle (1998b) to explain the replacement of many ancestral metabolic eukaryal proteins by bacterial ones in the course of eukaryal evolution. One could even wonder why this process has not completely confused the original relationships between DNA informational proteins (such as the orthology between some archaeal and eukaryal proteins). The answer might be that

### Table 1. Viral and plasmid proteins homologous to bacterial DNA replication proteins.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Species</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DnaB (E. coli)</td>
<td>Protein Ban, coliphage P1</td>
<td>Gene P protein, coliphage HK022</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Protein GP40, bacteriophage SPP1 (B. subtilis)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Protein GP12 coliphage P22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF3 plasmid (Chlamydiae sp.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF1, plasmid pCHL1 (Chlamydiae trachomatis)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF pscL (Streptomyces clavuligerus)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Protein RepB of bacteriophage MAV1 (Mycoplasma arthritidis)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proteins g4 of coliphages T3 and T7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proteins g61 α of coliphage T4</td>
</tr>
<tr>
<td>DnaC/DnaI/DnaA</td>
<td>ORF12, bacteriophage r1t of Lactococcus lactis</td>
<td>Gene Ntp, bacteriophage Φge1 of Lactococcus sp.</td>
</tr>
<tr>
<td>DnaG</td>
<td>Proteins G58 protein of mycobacteriophage</td>
<td>L5, gp61 of coliphage T4, α of coliphage P4, gp4 of coliphages T3 and T7</td>
</tr>
<tr>
<td>DnaE (E. coli)</td>
<td>YorL in phage-like element (Bacillus)</td>
<td></td>
</tr>
<tr>
<td>DnaD (B. subtilis)</td>
<td>ORF 48 of bacteriophage phi-PVL (iteration 1)</td>
<td></td>
</tr>
<tr>
<td>Protein Ssb</td>
<td>ORF 45, bacteriophage ΦPV (Staphylococcus aureus)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ORF 12, bacteriophage TP901-1, Lactococcus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gene 36 protein of bacteriophage Ssp1 (B. subtilis)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protein p116 of plasmid RK2 (E. coli)</td>
<td></td>
</tr>
<tr>
<td>Topo III</td>
<td>Gene TraE, plasmid RP4 (E. coli), Rf751 (Enterobacter)</td>
<td>Gene gamma, plasmid pDB101 (Streptococcus)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF00020, plasmid pMRC01 (Lactococcus)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Trs I, plasmid pMRC01, (Lactococcus lactis), plasmid pG01 (Staphylococcus)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pXO1 Bacillus antracis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>protein γ of pDB101 (Streptococcus pyogenes)</td>
</tr>
</tbody>
</table>

The non-redundant (nr) protein sequence database at the National Center for Biotechnology Information (NII, Bethesda, MD, USA) was searched for viral or plasmid homologues of *E. coli* and *B. subtilis* replicative proteins using the PSI-BLAST program (Altschul et al., 1997). When orthologues have different names in *E. coli* and *B. subtilis*, the name corresponds to *E. coli*. DnaC is only present in *E. coli*, whereas DnaI and DnaD are only present in *B. subtilis*. Proteins listed were retrieved either at the first step of the search or after one iterative search. The expect value for PSI-BLAST iteration was 0.001. In these conditions, no archaeal or eukaryotic homologues were retrieved, except in the case of Topo III. However, the scores were better for the plasmid-encoded Topo III listed here than for cellular proteins. DnaC, DnaI and DnaA are retrieved by successive iterations, in agreement with previous suggestions that these proteins are distantly related homologues (Koonin, 1992).
non-orthologous gene displacement was only rarely successful in such an integrated process as DNA replication.

Most interestingly, a well-documented case of non-orthologous displacement of a cellular protein by a viral or plasmid one has indeed been identified in the evolution of the transcriptional apparatus of mitochondria and chloroplasts (Gray and Lang, 1998). In most mitochondria, the original RNA polymerase of bacterial origin has been displaced by an RNA polymerase that does not belong to the family of cellular RNA polymerases (which are homologues in the three domains) but is instead homologous to RNA polymerases encoded by bacteriophage T3, T7 and SP6 and by several linear mitochondrial plasmids. The ancestral bacterial RNA polymerase is still present in the genome by several linear mitochondrial plasmids. The ancestral bacteriophage RNA polymerase is still present in the genome of the mitochondrion from the protist *Reclinomonas americana* (Lang et al., 1997), indicating that non-orthologous gene displacement occurred during the diversification of mitochondria inside the eukaryal lineage. In chloroplasts, part of the genome is transcribed by the original bacterial RNA polymerase, whereas another part is transcribed by a bacteriophage-like RNA polymerase closely related to its mitochondrial counterpart. Although the mitochondrial and chloroplast bacteriophage-like RNA polymerases are usually nuclear encoded, one of them is encoded by a gene located on a plasmid-like insert inside the mitochondrial genome of the brown alga *Pylaiella littoralis* (Rousvoal et al., 1998). The original non-orthologous replacement might thus have involved either a cryptic bacteriophage in the α-proteobacterial ancestor of mitochondria or a eukaryotic linear plasmid phylogenetically related to bacteriophage T3/T4/SP6.

Displacement of cellular proteins by viral or plasmid-encoded proteins could explain why the mechanisms for DNA synthesis and processing are less conserved than those for protein synthesis

The present hypothesis involving viruses and plasmids in the evolution of cellular DNA informational proteins readily explains why the mechanism for protein synthesis is much more conserved between the three domains of life than the mechanisms for DNA replication, repair, recombination and transcription. As already noticed, it has been proposed that the translation apparatus is highly similar between all extant cellular organisms, because it was already established and very efficient before the time of LUCA, whereas the mechanisms dealing with DNA were finalized independently in the three domains (Mushegian and Koonin, 1996; Olsen and Woese, 1997; Woese, 1998). I suggest instead that all mechanisms for information processing were fully operational at the time of LUCA, but that mechanisms for DNA replication, recombination, repair and transcription have been greatly modified by non-orthologous displacement from viruses or plasmids to cells. This could not have occurred for the translation apparatus, as viruses and plasmids cannot be the source of new genes involved in protein synthesis (with the exception of a few regulatory proteins).

Would it be possible to reconstruct the evolutionary history of the DNA replication apparatus?

DNA replication is a mechanism central to cellular life. Accordingly, the history of this mechanism is especially relevant to our understanding of early cellular evolution. The fact that most replicative proteins are non-homologues between bacteria and archaea/eukarya, whereas they are orthologues in all lineages of the same domain (e.g. all bacteria contain DnaA, DnaB, DnaG, DNA gyrase; all archaea contain an MCM-like protein, a eukaryal-like primase, at least one DNA polymerase B, Topo VI, etc.) suggests that a massive displacement occurred soon after the separation of archaea and bacteria. If the root of the universal tree is located in the bacterial branch, as is usually assumed (Olsen and Woese, 1997), this displacement could have occurred either in the bacterial or in the archaeal/eukaryal branch. However, if the root is in fact located in the eukaryotic branch, as suggested by new methods of phylogenetic reconstruction (Lopez et al., 1999b; Brinkmann and Philip, 1999), this event could only have occurred in the bacterial branch. It would be important to determine whether massive non-orthologous gene displacement actually occurred in the bacterial branch or in the archaeal one, as this would tell us which of the two replication systems, the bacterial or the archaeal/eukaryal ones, was originally present. However, this is a difficult task. The fact that the number of replicative proteins with a plasmid or viral homologues is much larger in bacteria than in archaea or eukarya could suggest that massive displacement of ancestral cellular genes by viral or plasmid-encoded ones occurred in bacteria. However, this observation might be skewed by the pattern of viral and plasmid sequences available. Indeed, there is presently only one complete sequence of an archaeal virus in the databases, and most eukaryotic viruses and plasmids are from animals or fungi. As viruses and plasmids are ubiquitous in the three domains (Zillig et al., 1996), it is a priori unclear why gene displacement from them would have been more important in a particular branch of the universal tree. A possibility is that a massive displacement occurred by chance either in archaea or in bacteria, giving the false impression that this lineage is especially prone to such an event. It is even plausible to imagine a single displacement event that involved several proteins at once interacting together and encoded by a single virus or plasmid.

A clue to the question of which lineage has been most affected by non-orthologous gene displacement might come from the observation that bacteria often use fewer...
polypeptides than archaea to perform the same function (Edgell and Doolittle, 1997). This suggests that bacteria have been especially affected by non-orthologous gene displacement, because such an event should a priori often be associated with streamlining (Forterre and Philippe, 1999). This happened, for example, when the four-subunit bacterial RNA polymerase was displaced by the monomeric bacteriophage RNA polymerase in the evolution of mitochondria and chloroplasts. However, as the precise mechanism of DNA replication in archaea is presently unknown, it remains to be seen whether bacteria indeed use fewer proteins than archaea for this particular process.

In any case, the present hypothesis about the role of non-orthologous gene displacement and of viruses and plasmids in the history of DNA informational proteins implies that reconstitution of this history will be very difficult. In particular, such displacement might have confused not only the phylogenies of the proteins directly affected, but also those of their interacting partners in macromolecular complexes. Indeed, the phylogeny of bona fide orthologues can be erroneous if their rates of evolution in one or several lineages has been modified because they interact physically with different protein partners (Dickerson, 1971). For example, if non-orthologous gene displacement indeed occurred more frequently in bacteria, this could explain why bacterial proteins have long branches in many universal protein trees (Forterre and Philippe, 1999).

The importance of viruses and plasmids in the history of life

The hypothesis proposed here implies that viruses and plasmids have played a much more important role in cellular evolution than is usually presumed, that is, not only as tools for genomic rearrangement, but also as a significant source of protein diversity. In particular, if some cellular proteins have been through a viral or plasmid stage during part of their evolutionary history, they might have benefited from a favourable environment to undergo refinement via rapid evolution in a novel context, possibly explaining why they were later able to displace some of their cellular functional relatives successfully.

After the Woesian revolution based on rRNA sequence comparison, viruses have been excluded from the universal tree of life. Recent studies in molecular evolution, enriched by the outcomes from cellular genome projects, have again focused attention on cellular evolution, viruses being left out. For these reasons, the study of viruses and plasmid evolution, as well as their roles in the history of cellular lineages, has been much neglected in recent times. As a consequence, the sampling of virus and plasmid sequences presently available is extremely skewed. This is a major difficulty in evaluating the validity of hypotheses such as those presented here. In my opinion, the isolation, characterization and sequencing of new viruses and large plasmids from the three domains of life and for all major branches in the three domains should be promoted more strongly in the future.

Acknowledgements

I am grateful to Dave Musgrave, Purificacion Lopez-Garcia, Bernard labedan and Marc Blight for critical reading of the manuscript and to an anonymous reviewer for important advice and notification of the DnaD bacteriophage homologue. Work in our laboratory is supported by grants from the Association de Recherche contre le Cancer and the European Community program Cell factory (BIO4-CT 96-0488).

References


Doolittle, W.F. (1998b) You are what you eat: a gene transfer


