Université Paris-Sud
Centre d'Orsay

THESE

Présentée pour obtenir le grade de

Docteur en Sciences de l'Université Paris-Sud

Par

Jonathan FILEE

Phylogénie moléculaire des gènes viraux impliqués dans le métabolisme et la réplication de l'ADN

Soutenue le 7 octobre 2002

Devant la commission composée de :

Pr Pierre CAPY	Examinateur
Dr Michael CHANDLER	Rapporteur
Dr Purificación LOPEZ-GARCIA	Rapporteur
Dr Jaqueline LAURENT	Examinateur
Dr David PRANGISHVILI	Examinateur

PREAMBULE

Cette thèse, initiée en stage de DEA sous la responsabilité de Jacqueline Laurent, a débuté en octobre 2001 dans le laboratoire de Patrick Forterre grâce à un financement de thèse du Ministère de la Recherche dans le cadre de l'Ecole Doctorale « Gène, Génome, Cellule » de l'Université Paris XI. Ces recherches ont bénéficié du soutien du « Programme de Recherche Fondamentale en Microbiologie et Maladies Infectieuses et Parasitaires » (PRFMMIP) du Ministère de la Recherche.

TABLE DES MATIERES

INTRODUCTION

L'introduction générale de ce manuscrit se compose naturellement de trois parties correspondant aux trois éléments du sujet de cette thèse.

En guise d'introduction il m'apparaît important de poser comme point de départ l'arbre phylogénétique (presque) universellement accepté du vivant proposé par Woese en 1987 (Woese 1987). Je parlerai ensuite de la question de l'enracinement de cet arbre, de la nature du dernier ancêtre commun universel (LUCA) et des incongruences des phylogénies moléculaires en fonction de la nature de la molécule étudiée. Cela me permettra de discuter du concept de « transfert horizontal de gène » et de l'importance supposée des éléments génétiques mobiles, dont les Virus, dans ce phénomène.

J'en viendrai ensuite à parler dans une seconde partie de ce qui me semble être les « oubliés » des Sciences de l'Evolution : les Virus. J'aborderai brièvement leur découverte et l'historique de leur étude, puis leur extraordinaire diversité. Je discuterai de l'état des connaissances concernant leurs modes d'Evolution et la relation évolutive hôte/parasite. J'en arriverai ainsi à poser la question de leurs impacts sur l'Evolution de leurs hôtes cellulaires.

Finalement dans une troisième partie je discuterai du rôle précédemment abordé des Virus dans un cadre particulier, celui de l'Evolution de l'appareil de réplication de l'ADN. Il sera question du pourquoi de l'étude de la réplication plutôt qu'un autre système, et en particulier des profondes différences entre le système de réplication des Bactéries, d'une part et des Archaea et Eucaryotes, d'autre part. J'en viendrai ensuite à expliciter les différentes hypothèses proposées pour rendre compte de cette observation et en particulier l'hypothèse fondatrice de cette thèse proposée par Forterre en 1999 (Forterre 1999) qui postule un remplacement massif de gènes cellulaires par des gènes viraux non-homologues dans les différentes lignées.

Commentaire : Si j'ai bien compris cette page présente le plan de l'introduction et non le plan du manuscrit. Je crains que le lecteur s'attende à autre chose : il faudrait donc le prévenir.

CHAPITRE A

L’arbre universel du vivant.

I. Brefs rappels historiques

On peut estimer que la première classification phylogénétique proposée pour représenter l'histoire de la vie, remonte à 1866 avec la publication par Ernst Haeckel de *Generelle Morphologie* (figure 1) (Haeckel 1866). En plus des 2 règnes traditionnellement reconnus (Plantae et Animalia), Haeckel y adjoint un troisième, Protista, regroupant la plupart des microorganismes.

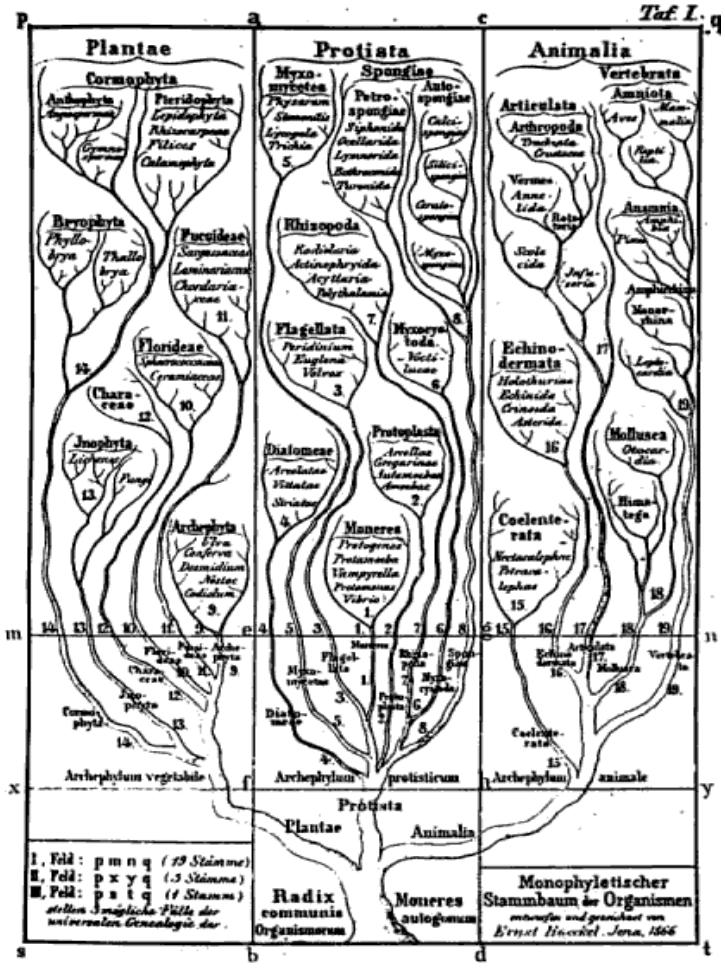

Figure 1 : Phylogénie extraite du Generelle Morphologie (Haeckel 1866).

Pour construire cet arbre, Haeckel se base sur l'existence des preuves matérielles de l'évolution : les fossiles. Cette phylogénie représente donc une généalogie des espèces, passant par différents "grades", pour finalement aboutir aux espèces actuelles. Malheureusement la pauvreté des archives paléontologiques limite considérablement la reconstruction des phylogénies avec cette méthode. Pendant plus d'un siècle les différentes phylogénies proposées par Haeckel ne vont pas, à quelques exceptions près, connaître de modifications. Il faut attendre les années 60 pour que la reconstruction phylogénétique connaisse de profonds bouleversements méthodologiques avec l'avènement de la systématique phylogénétique (Hennig 1966). Willy Hennig introduit une nouvelle façon de concevoir et d'utiliser le concept "d'homologie". L'homologie peut se définir comme les ressemblances existant entre différentes espèces en raison de leur ascendance commune. Un caractère homologue connaît deux états, un état "primitif" ou un état "dérivé" (figure 2). Avec l'approche Hennigienne (cladistique) c'est uniquement le partage d'un caractère "évolué" entre différentes espèces (ou synapomorphie) qui est signe d'une parenté étroite. Tout les taxas issus d'un ancêtre commun partageant une synapomorphie forme un groupe monophylétique (figure 2).

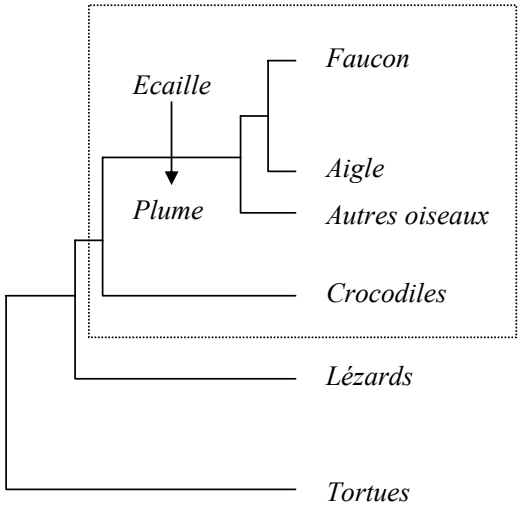

Figure 2 : Phylogénie des Reptiles et Oiseaux. Au sein du groupe indiqué en pointillé, la plume est un caractère dérivé (synapomorphie) et l'écaille un caractère primitif. Cette synapomorphie permet de former un groupe monophylétique regroupant tous les oiseaux.

Toutefois l'utilisation de caractères morphologiques ou biochimiques trouve rapidement ses limites quand il s'agit de comparer des organismes très divergents, surtout avec les microorganismes. Les naturalistes se heurtent à l'impossibilité de reconnaître des caractères

homologues entre tous ces organismes, ce qui entraîne d'importantes difficultés pour proposer un arbre phylogénétique universel du vivant. Un tournant a lieu en 1965 quand E. Zuckerland et L. Pauling (Zuckerland et Pauling 1965) ont l'idée d'utiliser pour la première fois des caractères moléculaires pour construire une phylogénie. En utilisant les séquences primaires d'une protéine, la chaîne β de l'hémoglobine, ceux-ci proposent la première phylogénie moléculaire des vertébrés qui est très similaire aux phylogénies obtenues avec des données morphologiques ou paléontologiques.

Commentaire : *Je préfère déjà comme cela, on verra pour le gloassaire, mais ce ne sera peut-être pa snécessaire*

En quelques années les phylogénies moléculaires basées sur différents gènes ou protéines se multiplient et prennent un essor considérable avec la mise au point des techniques rapides de séquençage de l'ADN (Sanger et al. 1977). La révolution apportée par les phylogénies moléculaires va connaître en 1977 un tournant décisif avec la publication de la première phylogénie moléculaire universelle du vivant par Carl Woese (Woese et Fox 1977). Les auteurs identifient un caractère homologue ubiquiste à tous les organismes : l'ARN ribosomique 16S/18S. Les auteurs sont à même de proposer une classification universelle du vivant en comparant les profils de migration sur gel d'électrophorèse bidimensionnelle des hydrolysats obtenus, à partir des ARN de différents organismes, en traitant cette molécule avec la ribonucléase T1. Avec le développement des techniques de séquençage de l'ADN, les régions du génome transcrites en ARN ribosomiques pourront ensuite être séquencées ce qui améliorera le degré de résolution des arbres (Woese 1987).

II. Les trois domaines du vivant

Les travaux de Woese sur l'ARN 16S/18S amènent une découverte surprenante qui va profondément bouleverser notre conception de l'origine et de l'évolution de la vie sur terre : celle de l'existence de deux lignées procaryotes totalement distinctes et non d'une seule comme on le pensait jusqu'alors (figure 3) (Woese et Fox 1977). Woese propose alors de définir deux nouveaux groupes :

- D'une part, les Eubacteria, qui regroupent la plupart des bactéries répertoriées jusqu'alors (cyanobactéries, bactéries à réaction de Gram positive ou négative).
- D'autre part, les Archaebacteria, qui regroupent principalement les bactéries méthanogènes. Elles représenteraient, selon Woese, le type procaryote ancestral du fait d'un métabolisme compatible avec l'atmosphère de la terre primitive (réduction anaérobie du dioxyde de carbone en méthane) .

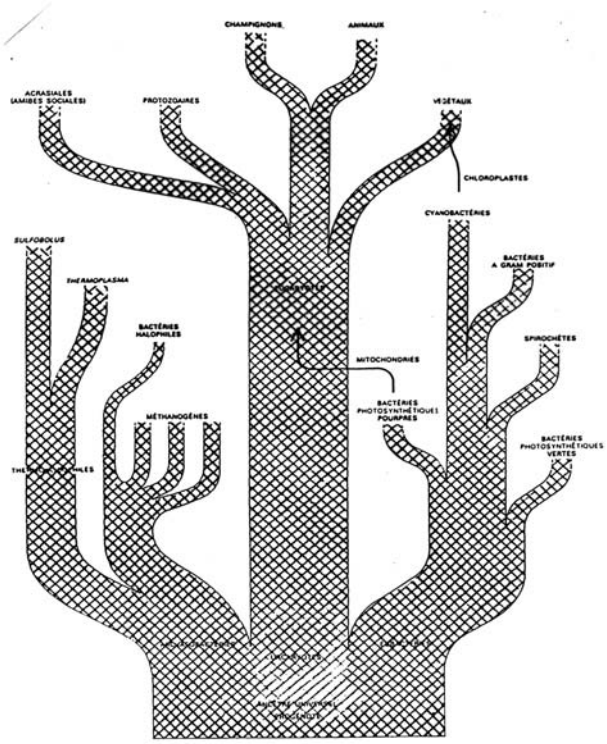

Figure 3 : L'arbre universel du vivant basé sur l'analyse des séquences de l'ARN ribosomique (Woese 1982).

Par la suite de nombreux autres procaryotes non méthanogènes se révéleront appartenir à ce groupe (Barns et al. 1996 ; Woese et Olsen 1986). Leur étude aboutira à la reconnaissance de 2 phyla majeurs au sein des Archebacteria : les Euryarcheota regroupant les halophiles, les méthanogènes, certaines thermoacidophiles et hyperthermophiles ; et les Crenarcheota regroupant surtout des thermoacidophiles et des hyperthermophiles. Le séquençage d'un nombre sans cesse croissant de séquences ribosomiques ne remettra pas en cause cette vision des choses (Figure 4). En outre, ces résultats seront confirmés par la plupart des phylogénies moléculaires basées sur des marqueurs protéiques (Matte-Taillez et al. 2002) .

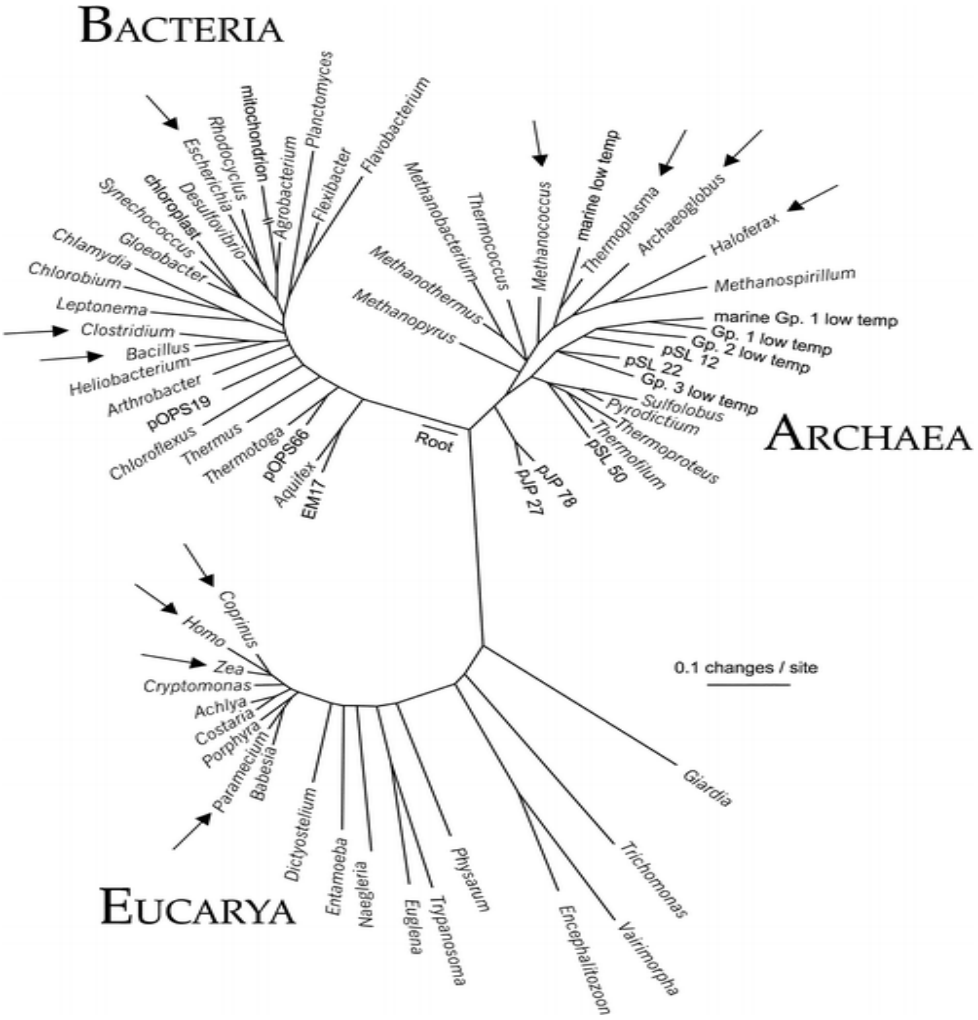

Figure 4 :L'arbre universel du vivant basé sur les séquences de l'ARN 16S/18S. La racine a été placée au sein des Bactéries (d'après Pace 1997)

Ces premières phylogénies moléculaires universelles vont également permettre de mettre un terme à un vieux débat concernant l'origine évolutive des organelles (mitochondrie et chloroplaste). Puisque les séquences d'ARN 16S des plastes et des mitochondries se placent dans l'arbre au sein des Eubacteria cela supporte très fortement l'idée d'une origine endosymbiotique pour ces organites (Eubacteria « phagocytée » par un Eucaryote primitif) (Margulis 1971). Les mitochondries dériveraient d'une α-Protéobactérie (Gray 1998) et les plastes d'une Cyanobactérie (Woese 1975). Un nombre croissant d'études basées sur des gènes nucléaires et mitochondriaux ont confirmé ces relations de parenté, étayant, en outre, l'idée d'un évènement unique d'endosymbiose d'une part pour la mitochondrie (Gray et al. 1999) et d'autre part pour le chloroplaste (Moreira et al. 2000). Toutefois, l'histoire évolutive de ces organites semble avoir suivi un chemin particulièrement chaotique avec de nombreux évènements d'acquisitions secondaires ou tertiaires d'un organite photosynthétique (phagocytose d'un eucaryote déjà pourvu de chloroplastes)(Moreira et Philippe 2001). De plus, de nombreux évènements de pertes secondaires de la mitochondrie ou du chloroplaste ont pu être mis en évidence. En effet, certains protistes amitochondriaux possèdent dans leurs génomes des gènes typiquement d'origine bactérienne et non Eucaryote, très probablement hérité de l'endosymbiose d'une organelle. Cette organelle ayant secondairement été perdues mais certains gènes seraient resté conservé. C'est par exemple le cas de la protéine HSP70 chez *Trichomonas vaginalis* (Germot et al. 1996) (Figure 5). Ce cas de figure a aussi été documenté pour la perte secondaire de la mitochondrie, par exemple chez les Microsporidies (Williams et al. 2002), ou du chloroplaste, par exemple chez les Trypanosomes (Hannaert et al. 2003).

Ainsi, la phylogénie moléculaire peut permettre, non seulement de retracer l'histoires des organismes, mais aussi d'inférer l'histoire évolutive de certaines caractéristiques biologiques, comme le devenir des organelles ou, comme nous allons le voir dans la prochaine partie, la nature du LUCA.

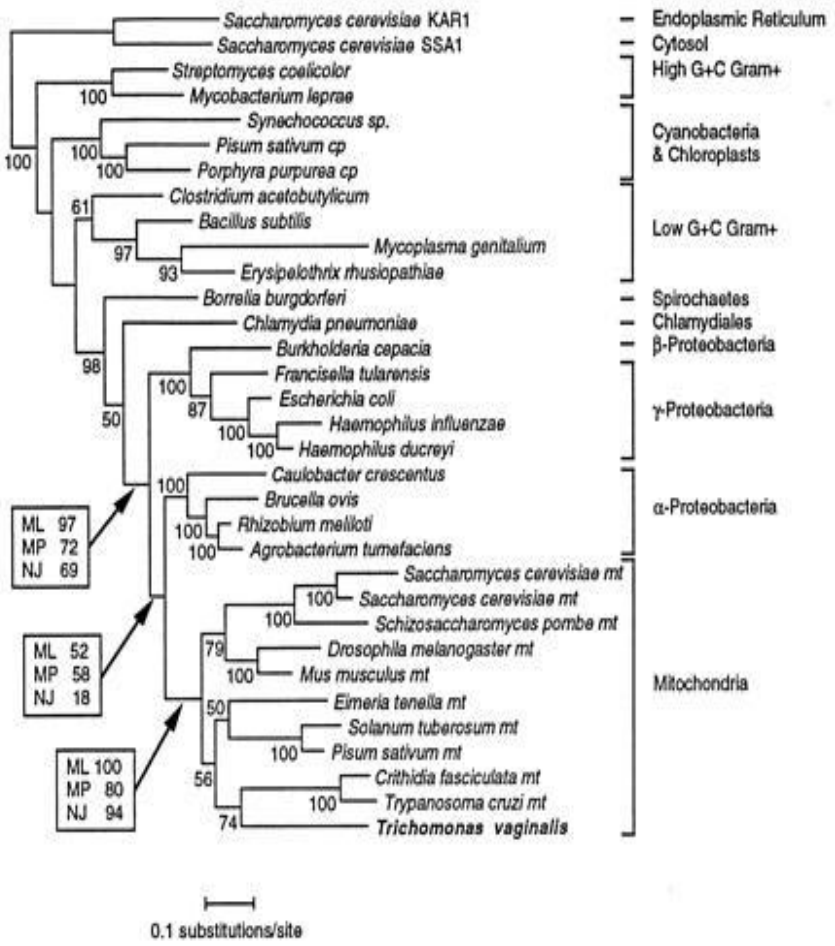

Figure 5 : Arbre phylogénétique de la protéine HSP70 chez les Bactéries et les organelles, raciné avec leurs homologues Eucaryotes. Le protiste amitochondrial *Trichomonas vaginalis* possède une HSP70 indiquée en gras qui se positionne au sein des HSP70 de mitochondries. Cette signature moléculaire indiquerait que ce protiste a possédé, puis perdu une mitochondrie au cours de son histoire évolutive. Tiré de Germot et al. 1996.

III. La racine de l'arbre universel et la nature du dernier ancêtre commun universel (LUCA)

L'arbre universel du vivant, basé sur les séquences d'ARNr 16S/18S, ne possède pas de groupe extérieur ; cet arbre est donc impossible à raciner, et l'ordre d'émergence des différents domaines ne peut donc pas être déterminé. Néanmoins, en se basant sur le fait que les Archéobactéries colonisent surtout les milieux extrêmes (en particulier les environnements très chauds dont certains auteurs estiment qu'ils correspondaient aux conditions qui prévalaient sur la terre primitive) de nombreux chercheurs, dont Carl Woese, ont pu penser que les Archéobacteries constitueraient le type ancestral : le LUCA serait donc procaryote et hyperthermophile.

a. L'enracinement de l'arbre universel

Pour contourner le problème de l'enracinement d'un arbre, Schwartz et Dayhoff (Schwartz et Dayhoff 1978) proposent d'utiliser des gènes déjà dupliqués (gènes paralogues) chez le dernier ancêtre commun à tous les organismes concernés, c'est à dire avant la spéciation ayant conduit à l'émergence de tous ces organismes. La phylogénie de l'un des gènes enracine alors la phylogénie de l'autre gène et *vice versa*. Cette technique sera mise en œuvre une décennie plus tard pour enraciner l'arbre universel du vivant avec les facteurs d'élongations EF-Tu/1α et EF-G/2 (Iwabe et al. 1989)(figure 6, avec une phylogénie plus récente de ces gènes) et les sous-unités α et β des ATPases V et F (Gogarten et al. 1989). Dans les 2 cas, la racine est positionnée dans la branche bactérienne, les Archéobactéries apparaissant plus proches des Eucaryotes, et les procaryotes étant donc bien polyphylétiques. Ces résultats seront généralement confirmés par les rares autres gènes a priori dupliqués chez l'ancêtre (Brown et Doolitle 1995). Pourtant ces résultats ne vont pas tarder à être fortement débattus.

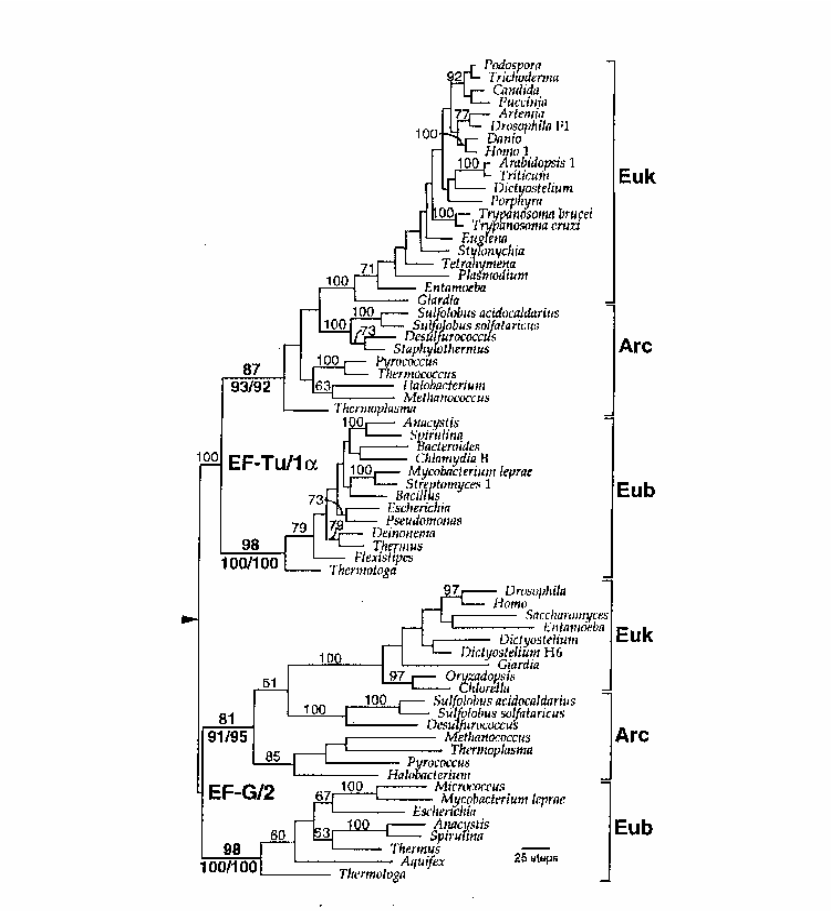

Figure 6 : Phylogénie basée sur les facteurs d'élongation EF-Tu/1α et EF-G2 enracinant l'arbre universel dans la branche des bactéries (extrait de Baldauf et al. 1996). Arc : Archéobactéries, Eub : Eubactérie, Euk : Eucaryote.

Tout d'abord, en ce qui concerne le choix des gènes dupliqués. Il est impératif que celui-ci n'ait été dupliqué qu'une seule fois avant le LUCA et que son histoire évolutive n'implique pas des transferts horizontaux, ce qui pourrait fausser l'interprétation des phylogénies. Or, concernant les ATPases, on a découvert l'existence, chez des bactéries, d'ATPases de type V, qui sont normalement présentes uniquement chez les Archébacteries et les Eucaryotes (Tsunami et al. 1991 ; Kakinuma et al. 1991) et celle d'une ATPase "bactérienne", de type F, chez une Archéobactérie (Sumi et al. 1992). On ne peut donc pas exclure soit qu'il se soit produit deux duplications successives pour les ATPases chez le LUCA (qui aurait donc

possédé non pas 2 mais au moins 4 gènes codant pour les sous-unités des ATPases) suivies de nombreuses pertes de gènes dans chaque lignées, soit plusieurs transfert horizontaux entre les 3 domaines du vivants (Hilario et Gogarten 1993). Ceci complique donc sérieusement l'interprétation de cet arbre au niveau de la position de la racine (Forterre et al. 1993). La seconde critique est que les paralogues ont divergé depuis si longtemps que les séquences des molécules ne sont alignables que sur de courts segments, ce qui conduit à des phylogénies basées sur un très faible nombre de sites, par conséquent de faible qualité au niveau du signal phylogénétique (Forterre et al. 1992, Philippe et Forterre 1999).
Pris ensemble, ces résultats suggèrent que le positionnement de la racine de l'arbre universel au sein des bactéries pourrait être artefactuel. La nature procaryote de LUCA est donc remise en question.

b. Le LUCA possédait il un noyau ?

Actuellement, le point de vue dominant postule que les procaryotes ont précédé les eucaryotes, ces derniers résultant de la complexification croissante d'une bactérie ou d'une Archéobactérie. Alternativement, de nombreux auteurs ont aussi proposé que les eucaryotes résultent de l'association entre une bactérie et une Archéobactérie soit par fusion (Zillig 1987 ; Golding et Gupta 1995) soit par symbiose métabolique (Martin et Muller 1998 ; Moreira et Lopez-Garcia 1998). Néanmoins, de nombreux exemples d'évolution par simplification sont également bien documentés, y compris pour des événement aussi importants que des pertes d'organelles (voir section II. Les trois domaines du vivant) ou des fortes réductions de la taille du génome et de son contenu en gène (par exemple chez les bactéries des genres Mycoplasma ou Rickettsia.). Ceci a conduit plusieurs auteurs à proposer que le LUCA ait été un eucaryote primitif et que les procaryotes résultent d'une évolution par simplification, notamment par perte du noyau (Reanney 1974 ; Doolitle 1978, Forterre et Philippe 1999).

c. Le LUCA était il hyperthermophile ?

A l'origine, les partisans d'un LUCA hyperthermophile se sont appuyés sur les résultats des phylogénies moléculaires. En effet dans les phylogénies obtenues à partir des séquences d'ARNr 16S/18S (comme pour celles obtenues avec certains marqueurs protéiques) les organismes hyperthermophiles occupent la base des groupes Archéobactéries et Bactéries (figure 6). Si ce placement reste peu discuté pour les Archéobactéries (Forterre et

al. 2002) il n'en va pas de même pour les Bactéries. En effet il été démontré que les organismes hyperthermophiles tendent à enrichir leurs ARN structuraux en nucléotides G et C par rapport aux organismes mésophiles. Un artefact dû à ce biais de composition qui différencie les organismes thermophiles des organismes mésophiles pourrait induire le regroupement d'organismes à taux de G-C équivalents dans les phylogénies (Galtier et Lobry 1997). De plus, d'après une analyse récente de jeux de données d'ARNr 16S chez les bactéries, le regroupement à la base des taxons hyperthermophiles résulterait surtout de la présence de sites évoluant rapidement (Brochier et Philippe 2002). En ne retenant que les sites évoluant plus lentement, considérés comme ayant conservé le « vrai » signal phylogénétique, ces auteurs placent à la base de l'arbre des Bactéries des organismes mésophiles appartenant aux groupes des Planctomycétales. Enfin, une topoisomérase particulière, la 'reverse gyrase', enzyme-clé nécessaire à l'adaptation à des environnements hyperthermophiles (Forterre 2002), semble avoir été acquise par les Bactéries postérieurement à la divergence des trois domaines, via des événements de transferts horizontaux en provenance d'Archéobactérie (figure 7, avec une phylogénie plus récente) (Forterre et al. 2000). Le LCA des Bactéries ne possédaient donc probablement pas de Reverse Gyrase et ne devaient donc pas être adapté à des niches écologiques très chaudes.

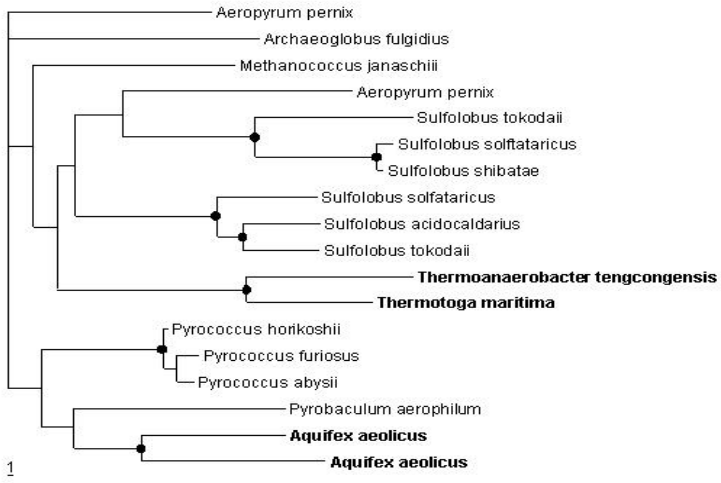

Figure 7 : Arbre non raciné de maximum de vraisemblance basé sur la protéine codant pour la Reverse Gyrase.

L'arbre est issu d'une recherche rapide en utilisant le programme PROTML (Adachi et Hasegawa 1996) et avec le modèle de substitution en acide aminé JTT-F. Les valeurs de bootstrap sont calculées en utilisant la méthode RELL appliquée aux 1000 meilleurs arbres. Les valeurs supérieures à 95% sont indiquées avec un rond noir. Les Bactéries sont indiquées en gras. La barre d'échelle représente le nombre de substitutions pour 100 sites par unité de longueur de branche.

La polyphylie des Bactéries thermophiles indique que le gène a probablement été acquis indépendamment, dans au moins deux lignées bactériennes différentes, via des transferts horizontaux en provenance des Archéobactéries (Forterre et al. 2000).

En somme, les interprétations de ces résultats convergent vers une sévère remise en cause de la position basale des Bactéries hyperthermophiles dans l'arbre ARNr 16S : on ne peut exclure une origine mésophile des Bactéries, secondairement adaptées à l'hyperthermophilie pour certaines d'entre elles.

Par ailleurs, à l'aide de méthodes de Maximum de Vraisemblance, basées sur un modèle de Markov, il a aussi été possible d'estimer le taux de G+C de l'ARNr du LUCA (Galtier et al. 1999, Galtier 2001). Sachant que le taux de G+C moyen des ARNr bactériens tourne autour de 50 à 55%, tandis que tous les ARNr des procaryotes hyperthermophiles ont des taux de GC supérieur à 60% (Galtier et Lobry 1997), il est donc possible d'établir une corrélation entre le

taux de G+C et un mode de vie (hyperthermophile ou non). Les taux de G+C du LUCA, obtenus par Galtier et collaborateurs varient en fonction de la méthode utilisée entre 54% et 55,5% pour l'ARNr 23S et entre 56,1% à 57,3% pour l'ARN 16S. Dans tous les cas, ce taux de G+C de LUCA est incompatible avec une vie à haute température. Toutefois le même jeu de données analysé avec une autre méthode basée sur le Maximum de Parcimonie conduit à des résultats inverses (ARNr 23S avec un taux de GC de 57,8% et ARN 16S avec un taux de G+C de 62,8%) (Di Giulio 2000). En outre, ce dernier auteur, mettant à profit le fait que les organismes hyperthermophiles enrichissent leurs protéines en acides aminés chargés (Cambillau et Claverie 2000), a développé une technique pour estimer le biais en acides aminés chargés chez le LUCA (Di Giulio 2003). Les résultats obtenus indiquent que le LUCA aurait été plutôt thermophile ou hyperthermophile.

Pris tous ensemble, ces résultats, souvent contradictoires, mettent en lumière le fait que la vision souvent communément admise d'une origine chaude de la vie et d'un LUCA hyperthermophile reste aujourd'hui encore vivement discutée. En outre la possibilité d'un LUCA mésophile ouvre de nouvelles perspectives concernant son éventuelle nature aérobie ou non et sur l'identité de son acide nucléique (ARN ou ADN).

d. Le LUCA et l'aérobiose

Du fait de l'extrêmement faible solubilité du dioxygène à haute température (au delà de 40°C il n'y a pratiquement plus de dioxygène en solution aqueuse), l'existence d'un LUCA à la fois hyperthermophile et aérobie semble improbable. Mais si le LUCA était mésophile ou modérément thermophile, cela ouvre la possibilité de la présence d'un métabolisme ancestral aérobie. Cette hypothèse a été largement développée par José Castresana qui affirme que le LUCA était aérobie en s'appuyant sur des arguments génomiques et phylogénétiques (Castresana et Moreira 1999 ; Castresana 2001). Une part essentielle de l'argumentation est basée sur des enzymes clés de la respiration aérobie : les cytochromes oxydases. Il existe deux familles non-homologues de cytochrome oxydase : la famille « SoxM/SoxB » et la famille « bd ». Etant donné la large répartition phylogénétique des deux familles chez les bactéries, il est très probable que le dernier ancêtre commun (LCA) des bactéries possédait les deux enzymes et possédait donc un métabolisme aérobie. Les eucaryotes auraient acquis un

métabolisme aérobie avec l'endosymbiose de la mitochondrie, il n'y a donc pas de raison objective de penser que le LCA des eucaryotes était aérobie avant l'endosymbiose de la mitochondrie. La situation des Archéobactéries est particulièrement intéressante. En ce qui concerne la phylogénie des cytochromes oxydases de la famille SoxM/SoxB les Archéobactéries apparaissent polyphylétiques, à la fois pour le gène de type SoxM et pour le gène de type SoxB (Castresana 2001 ; Filée, résultats non présentés). L'analyse phylogénétique des gènes de la famille bd met en évidence une polyphylie marquée des Archéobactéries (Figure 8) (Filée et Myllykallio, résultat non publié). Au moins pour la famille "bd", il est donc fort probable que les Archéobactéries aient acquis le gène indépendamment les unes des autres via des transferts horizontaux en provenance de Bactéries. Ces observations supposent donc des histoires évolutives complexes pour les cytochromes oxydases des Archéobactéries, impliquant :

- soit un ancêtre aérobie dont les descendants auraient connu des duplications de gènes puis de multiples pertes sélectives.
- soit un ancêtre anaérobie, quelques descendants auraient bénéficié de transferts horizontaux de gènes en provenance des Bactéries pour leur permettre la vie aérobie. Cette hypothèse est aussi supportée par la probable origine hyperthermophile du LCA des Archéobactéries (Forterre et al. 2002)

Prises ensemble, ces constatations indiquent que, à la différence du LCA des Bactéries, le LCA des Eucaryotes et le LCA des Archéobactéries ne possédaient probablement pas de métabolisme aérobie. Indépendamment de la position de la racine de l'arbre du vivant, la présence d'un métabolisme aérobie chez le LUCA proposée par Castresana (Castresana 2001) reste donc très spéculative.

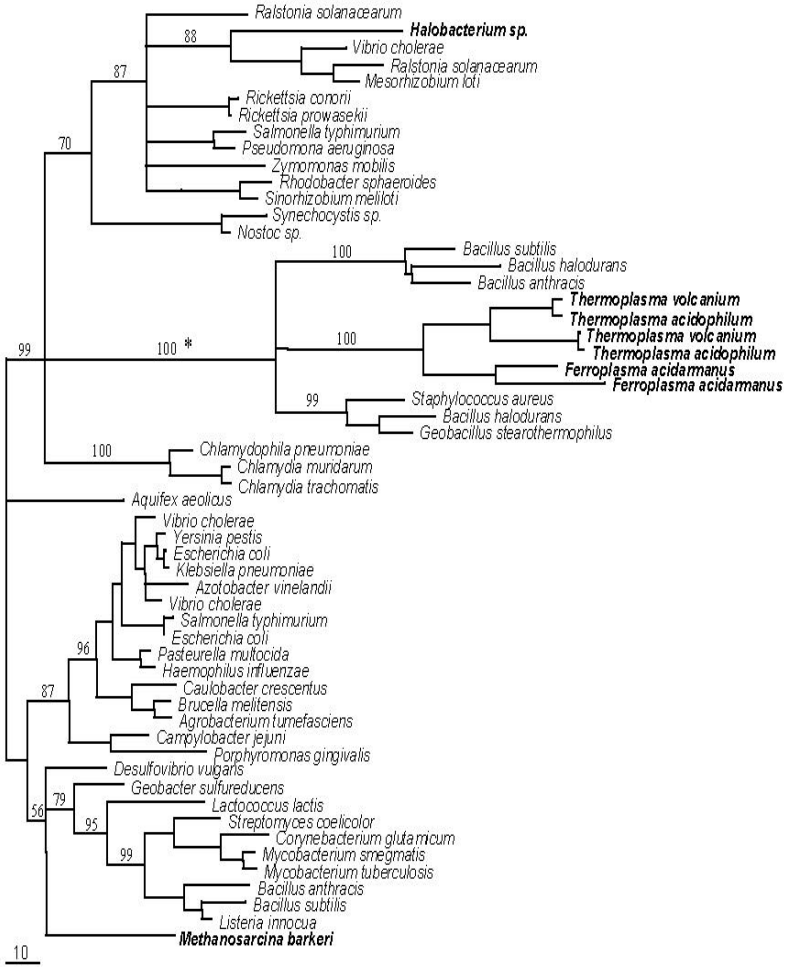

Figure 8 : Arbre de maximum de vraisemblance basé sur la protéine codant pour la Cytochrome oxydase de la famille « bd ».
L'arbre est issu d'une recherche rapide en utilisant le programme PROTML (Adachi et Hasegawa 1996) et avec le modèle de substitution en acide aminé JTT-F. Les valeurs de bootstrap sont calculées en utilisant la méthode RELL appliquée aux 1000 meilleurs arbres. Les Archéobactéries sont indiquées en gras. La barre d'échelle représente le nombre de substitutions pour 100 sites par unité de longueur de branche.
L'astérisque indique que le groupement est aussi supporté par plusieurs INDEL(s) au niveau de l'alignement des séquences.

Figure 6 : Arbre de maximum de vraisemblance basé sur la protéine codant pour la Cytochrome oxydase de la famille « bd ».
L'arbre est issu d'une recherche rapide en utilisant le programme PROTML (Adachi et Hasegawa 1996) et avec le modèle de substitution en acide aminé JTT-F. Les valeurs de bootstrap sont calculées en utilisant la méthode RELL appliquée aux 1000 meilleurs arbres. Les Archéobactéries sont indiquées en gras. La barre d'échelle représente le nombre de substitutions pour 100 sites par unité de longueur de branche.
L'astérisque indique que le groupement est aussi supporté par plusieurs INDEL(s) au niveau de l'alignement des séquences.

(Les arbres phylogénétiques résultats non publiés)

e. La nature du génome du LUCA

Si l'idée qu'un monde ARN aurait précédé notre monde ADN actuel est une vision aujourd'hui bien acceptée, il est difficile d'imaginer un LUCA hyperthermophile ayant possédé un génome ARN. En effet l'ARN est une molécule très peu thermostable qui s'hydrolyse très rapidement à des températures voisines de 100°C. De plus, la vitesse de dégradation de l'ARN à haute température est accélérée en présence de magnésium, un cofacteur essentiel de la plupart des réactions catalysées par les ribozymes.

La question de l'origine des génomes à ADN est indissociable de celle de l'origine et de l'évolution des enzymes informationnelles, et singulièrement des enzymes de la synthèse et de la réplication de l'ADN. Ces points seront largement développés dans la suite de ce mémoire. Dès à présent, d'une manière synthétique, on peut dire que les molécules informationnelles se ressemblent beaucoup plus entre Eucaryotes et Archéobactéries qu'entre Bactéries et l'un ou l'autre de ces deux groupes. Ceci est valable aussi bien pour les molécules universellement conservées (Olsen et al. 1997) qu'entre les molécules qui n'ont pas d'homologue au sein d'un des trois domaines [ceci étant particulièrement vrai pour les enzymes de la réplication (Mushegian et al. 1996 ; Olsen et al. 1997, Myllykallio et al. 2000)]. Ces observations ont été souvent interprétées comme la preuve que le LUCA possédait un génome ARN et que l'ADN avait été « inventé » indépendamment plusieurs fois (Mushegian et al. 1996, Leipe et al. 1999). Cette hypothèse, ainsi que les hypothèses concurrentes, seront discutées dans le chapitre C.

f. Conclusion

En l'état de nos connaissances, il est tout à fait impossible de dresser un portrait robot du LUCA. En effet, depuis les évènements de spéciations ayant conduit aux trois domaines du vivant, il y a plusieurs centaines de millions d'années, les organismes actuels ont beaucoup évolué et divergé, atténuant ou effaçant même les traces ancestrales.

D'un point de vue phylogénétique, les incongruences entre différents gènes compliquent sérieusement l'établissement de ce portrait robot. De nombreux obstacles empêchent en effet de reconstruire une phylogénie universelle du vivant, les principaux seront abordés dans la section V de ce chapitre.

D'un point de vue génomique, les analyses actuelles sont fortement dépendantes de l'échantillonnage utilisé, et l'afflux de données issues du séquençage de génomes complets

modifie régulièrement notre vision des choses. Le problème est d'autant plus crucial que notre connaissance de la biodiversité microbienne reste très parcellaire, au point que certains auteurs estiment que l'on sait cultiver moins de 1% de la biodiversité microbienne (Pace 1997). Il ne fait donc aucun doute qu'une meilleure connaissance (au niveau génomique en particulier) de cette extraordinaire diversité pourrait apporter des éclairages nouveaux sur les questions abordées ici.

IV. Biodiversité microbienne et écologie moléculaire.

Le clonage et le séquençage de l'ARN 16S par PCR directement à partir d'échantillons environnementaux a rendu possible l'étude de la biodiversité microbienne sans passer par la très contraignante étape de mise en culture (Amann et al. 1995). Cette technique appliquée à l'étude d'environnements très divers, depuis les sources hydrothermales jusqu'au tube digestif des mammifères, a permis la découverte d'un très grand nombre d'espèces nouvelles à la fois chez les Bactéries, les Archéobactéries et les Eucaryotes (figure 9).

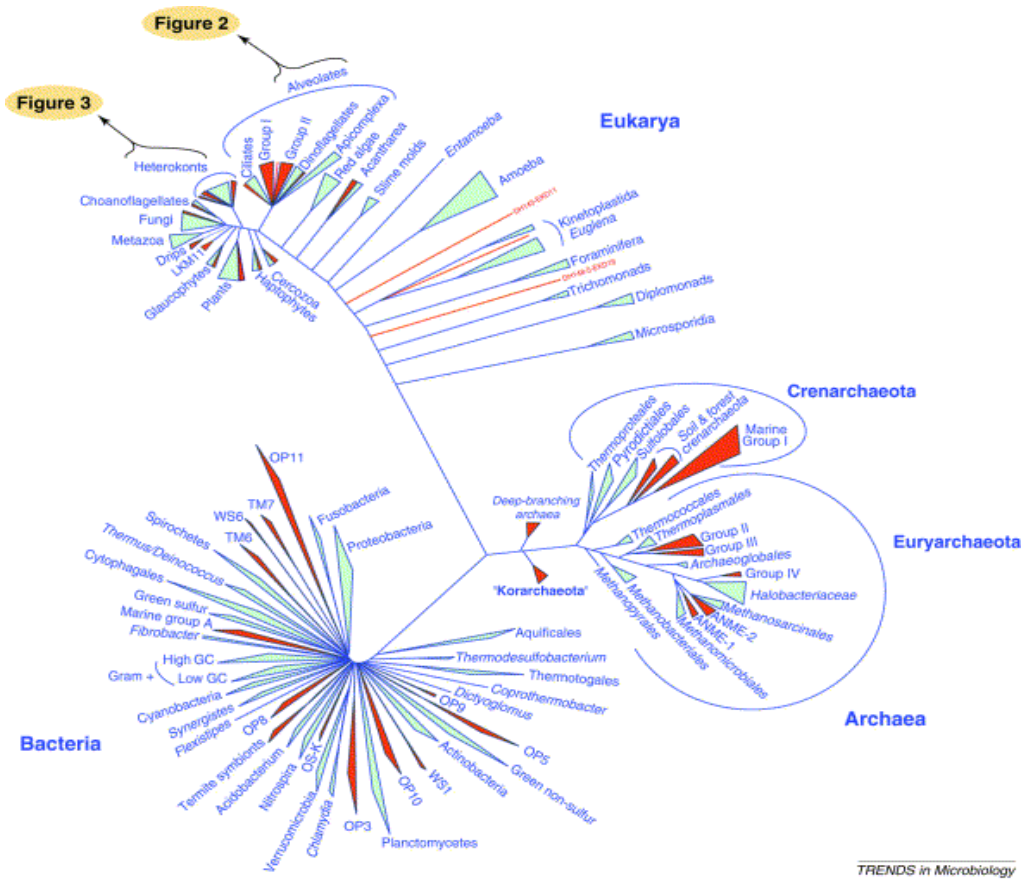

Figure 9 : Représentation schématique de l'arbre universel du vivant à la lumière des séquences d'ARNr 16S issues de l'environnement, les groupes indiqués en rouge ne sont connus que par leurs séquences d'ARNr 16S (extrait de Moreira et Lopez-Garcia 2002).

Chez les Bactéries, le nombre de phyla reconnus a presque doublé, et de nombreux phylums (ou divisions) tout entiers ne sont connus que par leurs séquences d'ARNr 16S comme le groupe WS6 (Dojka et al. 2000) ou le groupe OP11 (Dojka et al. 1998). Ces 2 nouveaux groupes semblent coloniser des environnements très différents (thermophiles et mésophiles en particulier, mais essentiellement dans des environnements anaérobies), ce qui indique potentiellement une vaste gamme de capacités métaboliques. De plus le degré de divergence des séquences d'ARN 16S à l'intérieur de ces 2 groupes est le plus élevé connu à ce jour chez les Bactéries (figure 10) (Dojka et al. 2000). Certains de ces phyla non-cultivés constituent donc des groupes très importants d'un point de vue évolutif, qui pourraient nous aider à mieux résoudre la phylogénie des Bactéries et en particulier à mieux caractériser leur LCA.
Chez les Archéobactéries, de nombreux nouveaux taxa appartenant aux 2 divisions traditionnellement reconnues (Euryarchaeota et Crenoarchaeota) ont été identifiés. De plus, deux nouvelles divisions d'Archéobactéries hyperthermophiles ont été proposées : un groupe basal les « Korarcheota » (Barns et al. 1996) d'une part et d'autre part les « Nanoarchaeota » (Huber et al. 2002) qui présente la particularité d'avoir une taille de génome et de cellule très réduites. Pour ce qui concerne les Archéobactéries, les taxons dans leur majorité sont à ce jour non-cultivables et notre aperçu de la biodiversité de ce domaine demeure donc très biaisé (Forterre et al. 2002). Il n'est donc pas impossible que nos connaissances sur l'évolution de ce domaine soient largement remises en cause dans les années qui viennent.

Enfin chez les Eucaryotes, de récentes études ont montré que la biodiversité en protistes était beaucoup plus importante qu'on ne le pensait (Diez et al. 2001 ; Lopez-Garcia et al. 2001 ; Moon-Van der Staay et al. 2001). La plupart de ces Eucaryotes microbiens ont été découverts dans des échantillons océaniques et semblent avoir une répartition géographique très large. Il est toutefois probable qu'une vaste diversité de protistes de petites tailles reste à découvrir dans d'autres environnements (eau douce, biotope terrestre incluant des milieux extrêmes) et qu'une meilleure connaissance de ces organismes constitue potentiellement une des clés pour résoudre les nœuds les plus basaux de la phylogénie des Eucaryotes (Moreira et Lopez-Garcia 2002).

Pris ensemble ces résultats indiquent que beaucoup d'organismes présentant un grand intérêt évolutif ne sont pas encore bien connus et caractérisés. Ils constituent en biologie évolutive un des domaines majeurs pour de futures recherches et on peut en espérer de nombreux apports concernant les questions abordées dans cette introduction.

Division	*Typical % difference*
Cyanobacteria	*13*
Fusobacteria	*13*
Termite group 1	*15*
TM7	*16*
OP10	*17*
Thermus/Deinococcus	*17*
Actinobacteria (high G+C gram positive)	*18*
Nitrospira	*19*
Thermotogales	*19*
Acidobacteria	*20*
Green sulfur	*20*
Verrucomicrobia	*20*
Cytophagales	*22*
Spirochetes	*22*
Green nonsulfur	*23*
Planctomycetes	*23*
Proteobacteria *(α-ε)*	*23*
Low G+C gram positive	*24*
WS6	*26*
OP11	*33*

Figure 10 : Taux de divergence maximum de l'ARNr 16S en % au sein de différentes lignées de Bactéries. (Tiré de Dojka et al. 2000).

V. Les incongruences entre phylogénies et l'importance des transferts horizontaux de gènes.

Le séquençage d'un nombre sans cesse croissant de gènes et leurs analyses phylogénétiques montre que les phylogénies de protéines sont très majoritairement en contradiction avec la phylogénie de « référence » basée sur l'ARNr 16S/18S. Ces incongruences peuvent être expliquées à la fois :

- par des artefacts de reconstruction phylogénétique tels que des différences de vitesse d'évolution entre séquences.
- par des processus biologiques tels que les transferts horizontaux de gènes ou encore la duplication d'un gène et des pertes indépendantes d'un des paralogues dans telle ou telle lignée.

a. Les artefacts de reconstructions phylogénétiques :

Les vitesses de fixation des mutations sur une séquence ne sont pas constantes entre différentes espèces pour un gène donné. Il en résulte une accentuation des inégalités des longueurs de branches, au-delà de la seule distance évolutive, dans les arbres phylogénétiques. Ces différences de vitesses d'évolutions entre les séquences peuvent générer des artefacts d'attraction des longues branches (ALB)(Felsenstein 1978)(figure 11).

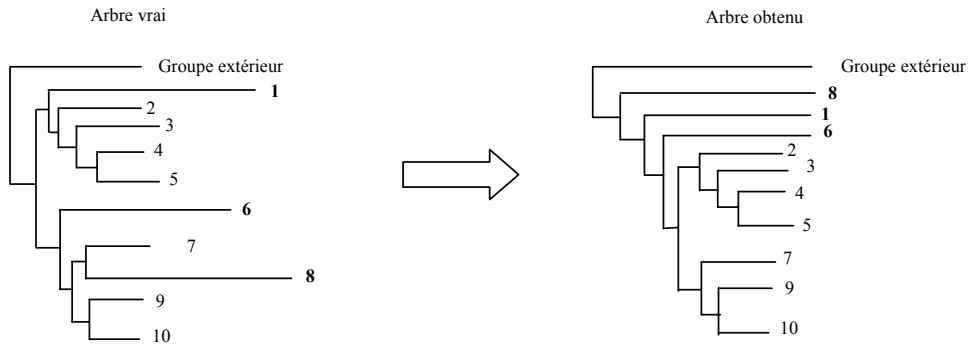

Figure 11 : Artefact d'attraction des longues branches

Dans l'exemple présenté, 3 taxons indiqués en gras évoluent plus vite que les autres, si le groupe extérieur utilisé est distant, les taxas évoluant vite vont se regrouper ensemble à la base de l'arbre formant une base asymétrique. D'après Philippe et Laurent (1998).

Indépendamment des véritables relations de parenté, si certaines séquences évoluent beaucoup plus vite que d'autres, celles qui évoluent moins vite gardent plus de caractères ancestraux et finissent par se ressembler plus entre elles qu'aux autres. Avec la plupart des méthodes « classiques » de phylogénie, les séquences évoluant rapidement se regroupent ainsi artificiellement ensemble (Philippe et Laurent 1998). Si on utilise un groupe extérieur éloigné, ce qui est généralement le cas pour les phylogénies retraçant des évènements anciens, on peut obtenir des arbres à base asymétrique où la longue branche du groupe extérieur « attire » les autres taxa à longue branche (figure 11). Par exemple, cet artefact semble être responsable du placement de la racine de l'arbre universel au sein des bactéries quand on utilise les paralogues des facteurs d'élongation (voir section III), ou du placement de certains protistes amitochondriaux à la base de l'arbre ARNr 18S des Eucaryotes (Philippe et Adoutte 1998). L'utilisation d'un groupe extérieur à branche courte est un bon moyen pour limiter les effets de l'attraction des longues branches.

Les différences observées entre les arbres utilisant des marqueurs différents peuvent aussi s'expliquer par la saturation mutationnelle des séquences. En effet, au cours du temps de nouvelles substitutions se produisent à des positions ayant déjà muté. Si l'on cherche à reconstruire l'histoire d'évènements anciens, le signal phylogénétique ancien peut être en partie remplacé par du signal phylogénétique plus récent qui n'est pas informatif pour le problème qui nous concerne (Philippe et Laurent 1998); d'où des différences dans les topologies des arbres.

b. Les transferts horizontaux de gènes.

Le fait que des organismes asexués comme les Bactéries puissent échanger du matériel génétique est un phénomène connu depuis presque 3/4 de siècle. En effet dès 1928, Griffith démontre que des souches non virulentes de pneumocoques peuvent, sous certaines conditions, se transformer en souches virulentes si l'on les met en contact avec une solution de bactéries virulentes détruites par la chaleur. Seize ans plus tard la nature chimique de l'agent responsable de la transformation sera mise en évidence : c'est l'ADN (Avery et al. 1944).

Le développement des techniques de séquençage de l'ADN et l'exploitation phylogénétique des données de la génomique ont considérablement popularisé l'idée selon laquelle les transferts horizontaux de gènes pouvaient constituer une des forces majeures de l'évolution

des procaryotes (Doolitle 1999a, Doolitle 1999b). La littérature fourmille d'exemples et de revues faisant la synthèse des gènes potentiellement affectés par les transferts horizontaux (pour revue voir Ochman et al. 2000 ; Koonin et al. 2001). Ces derniers sont considérés comme des facteurs évolutifs essentiels pour l'adaptation à de nouveaux environnements (nouvelle propriété métabolique, résistance aux antibiotiques) et certains auteurs pensent même qu'il pourrait jouer un rôle important dans la spéciation des organismes (Lawrence 1997).

A l'heure actuelle, la détection de transferts horizontaux procède par quatre approches différentes qui peuvent être complémentaires :

- Une approche intrinsèque qui se base sur l'hypothèse selon laquelle l'usage des codons et le contenu en GC constituent des signatures distinctes de chaque génome (Grantham et al. 1980). Ainsi les gènes présentant une composition nucléotidique ou un usage du code significativement différent de la moyenne pour un génome donné sont considérés comme potentiellement issus d'un transfert horizontal (Medigue et al. 1991 ; Lawrence et Ochman 1997 ; Garcia-Vallve et al. 2000). Cette méthode permet la détection de transferts récents car les fragments intégrés perdent progressivement les caractéristiques du génome donneur en prenant les caractéristiques du génome receveur.
- Une approche basée sur les similarités de séquences qui, à l'aide de programmes d'alignement global, de type BLAST (Altschul et al. 1997) permet de détecter au sein d'une famille homologue, un gène qui ressemble très peu à celui des taxons dont il est a priori évolutivement proche. Cette technique permet de détecter surtout des transferts entre organismes très éloignés, comme c'est le cas entre les 3 domaines [par exemple les transferts « massifs » de gènes détectés entre bactéries hyperthermophiles et Archéobactéries (Aravind et al. 1999 ; Nelson et al. 1999)].
- Une approche génomique qui, en comparant la présence et l'absence de gènes orthologues principalement chez des espèces/souches proches, permet de quantifier et de qualifier les flux de perte/acquisition de gènes (Lan et Reeves 1996 ; Hayashi et al. 2001).
- Enfin l'approche phylogénétique permet généralement *in fine* de valider les résultats obtenus par des approches plus globales, elle a l'inconvénient d'être lourde à mettre en œuvre, mais présente l'avantage de pouvoir parfois polariser la direction du transfert.

La question qui s'est rapidement posée fut de savoir si il existait des catégories de gènes plus sujets aux transferts que d'autres. Il est concevable d'imaginer que les gènes les plus couramment transférés sont ceux qui apportent un avantage sélectif. En effet les gènes de résistance à des antibiotiques ou codant pour des facteurs de virulences sont fortement sujets à transfert. Toutefois un très grand nombre de gènes identifiés comme potentiellement issus de transferts ne donnent a priori pas d'avantage sélectif important. Et dans le cas de l'acquisition d'un gène dont un homologue existe déjà au sein du génome, on voit mal comment le remplacement homologue du gène initialement présent procurerait un quelconque avantage sélectif. De plus, une part importante des gènes identifiés comme issus de transferts ne codent pas pour des fonctions connues (figure 12) (Garcia-Vallve et al. 2000). La présence de tous ces gènes potentiellement issus de transferts dans les génomes reste donc mal expliquée, et il serait particulièrement intéressant de quantifier ce flux d'acquisition de nouveau gène dans les génomes.

L'hypothèse selon laquelle les gènes « opérationnels » sont plus fréquemment transférés que les gènes « informationnels » a été plusieurs fois suggérée (Jain et al. 1999 ; Subramanian et al. 2000) mais pas systématiquement démontrée (Nesbo et al. 2001). Les protéines informationnelles seraient plus souvent associées à de grands complexes protéiques. Et selon les auteurs de cette hypothèse plus le produit d'un gène donné est impliqué dans des complexes avec beaucoup d'interactions entre protéines, moins le gène correspondant a de chance d'être, sinon transféré du moins conservé après un éventuel transfert. Cette vison est assez subjective, car de nombreux cas bien documentés de transferts horizontaux impliquant des gènes codant pour des protéines de la traduction [(par exemple pour la protéine ribosomale Rps14 chez certaines Bactéries (Brochier et al. 2000)] ou de la réplication [par exemple l'ADN polymérase de la famille B des Gamma Protéobactéries (Edgell et al. 1998)] ont été mis en évidence et le phénomène ne semble pas du tout négligeable au plan quantitatif (figure 12).

D'un point de vue biologique, les mécanismes d'acquisition d'un gène exogène sont bien connus (Zgur-Bertok 1999). En effet les processus de transduction, transformation ou conjugaison sont connus depuis les années 50. Dans ces processus, les éléments génétiques mobiles que sont virus, plasmides et transposons jouent un rôle clé. Pourtant le rôle de ces éléments, et singulièrement des virus, dans l'évolution des cellules est généralement sous-estimé (Balter 2000). Ne possédant pas d'ARN ribosomique ils ne trouvent pas leur place

dans l'arbre universel du vivant. Pourtant leurs acides nucléiques évoluent et de nombreux auteurs estiment leurs origines très anciennes (Forterre 1992 ; Hendrix et al. 1999 ; Benson et al. 1999 ; Balter 2000). L'évolution des virus et les rôles réciproques joués par les virus et leurs hôtes constitue donc un domaine de recherche encore peu développé, mais qui pourrait apporter des éclairages nouveaux sur une partie des questions abordées dans ce chapitre.

Organism	***HGT***	***Info. (%)***	***Cell (%)***	***Meta. (%)***	***Poor (%)***	***−(%)***
Archaeoglobus fulgidus	*179*	*6 (2.2)*	*22 (8.1)*	*17 (2.7)*	*31 (6.0)*	*103 (14.5)*
Aquifex aeolicus	*72*	*7 (3.2)*	*12 (3.8)*	*15 (3.6)*	*20 (6.4)*	*18 (6.9)*
Borrelia burgdorferi	*12*	*2 (1.1)*	*3 (1.7)*	*1 (0.8)*	*3 (2.2)*	*3 (1.3)*
Bacillus subtilis	*537*	*54 (11.5)*	*34 (5.8)*	*76 (8.0)*	*42 (7.3)*	*331 (21.8)*
Chlamydia pneumoniae	*55*	*6 (3.4)*	*4 (3.0)*	*10 (4.9)*	*6 (5.3)*	*29 (6.8)*
Chlamydia trachomatis	*36*	*5 (2.9)*	*5 (3.8)*	*9 (4.8)*	*0 (0.0)*	*17 (5.7)*
Escherichia coli	*381*	*23 (4.8)*	*41 (6.6)*	*42 (3.8)*	*27 (5.4)*	*248 (15.7)*
Haemophilus influenzae	*96*	*3 (1.1)*	*19 (7.0)*	*10 (2.2)*	*3 (1.4)*	*61 (12.0)*
Helicobacter pylori 26695	*89*	*11 (5.0)*	*3 (1.1)*	*5 (1.6)*	*3 (1.6)*	*67 (11.6)*
Helicobacter pylori J99	*80*	*7 (3.2)*	*6 (2.3)*	*5 (1.6)*	*4 (2.2)*	*58 (11.5)*
Mycoplasma genitalium	*67*	*26 (19.1)*	*10 (16.7)*	*11 (12.6)*	*6 (9.1)*	*14 (10.7)*
Methanococcus jannaschii	*77*	*8 (3.6)*	*11 (7.5)*	*6 (1.6)*	*7 (1.8)*	*45 (7.8)*
Mycoplasma pneumoniae	*39*	*8 (5.1)*	*0 (0.0)*	*3 (2.7)*	*2 (2.7)*	*26 (9.6)*
Methanobacterium thermoautotrophicum	*179*	*4 (1.7)*	*19 (8.8)*	*26 (5.6)*	*44 (11.1)*	*86 (15.5)*
Mycobacterium tuberculosis	*187*	*6 (1.6)*	*20 (4.9)*	*35 (4.0)*	*17 (3.2)*	*109 (6.2)*
Pyrococcus horikoshii	*154*	*10 (4.2)*	*11 (6.1)*	*8 (2.1)*	*23 (4.9)*	*102 (12.7)*
Rickettsia prowazekii	*28*	*8 (4.4)*	*3 (1.9)*	*12 (6.6)*	*4 (3.6)*	*1 (0.5)*
Synechocystis PCC6803	*219*	*14 (5.3)*	*31 (5.2)*	*15 (2.6)*	*22 (5.0)*	*137 (10.6)*
Thermotoga maritima	*198*	*13 (5.3)*	*18 (6.5)*	*55 (10.7)*	*47 (12.0)*	*65 (15.6)*
Treponema pallidum	*77*	*8 (4.5)*	*17 (8.7)*	*2 (1.4)*	*17 (10.6)*	*33 (9.2)*

Figure 12 : Classification fonctionnelle des gènes détectés comme acquis récemment par transfert horizontal (HGT) dans différents génomes issus de la banque de données COG (Tatusov et al. 2000). Les chiffres indiquent le nombres et le pourcentage de gènes ayant été acquis par HGT pour chaque taxon. Les groupes fonctionnels sont définis comme suit : Info, informationnelle ; Cell, processus cellulaire ; Meta, métabolisme ; Poor : faiblement caractérisé ; -, non présent dans les COG. Extrait de Garcia-Vallve et al. 2000.

CHAPITRE B

Virus et Evolution

I. La découverte des virus

Les premières études faisant état des virus datent de la fin du XIX^{e} siècle. En 1886, Mayer décrit l'existence d'un principe infectieux thermosensible responsable de la mosaïque du tabac. La première description de ce virus est publiée par Beijerinck en 1898 : un microbe capable de diffuser à travers de l'agar en raison de sa faible taille, pouvant être conservé pendant deux ans dans des feuilles de tabac séchées sans perdre sa pathogénicité et capable de pénétrer les cellules et de s'y multiplier. La même année Löffler et Frosch, grâce à leurs travaux sur la fièvre aphteuse, démontrèrent l'existence chez les mammifères d'agents pathogènes possédant les mêmes caractéristiques.

Les virus de bactéries ou bactériophages seront découverts simultanément par Twort (1915), sur des colonies de *Micrococcus,* et par d'Herelle (1917) sur des cultures de *Shigella.* Mais leur nature virale sera très débattue jusqu'au début des années 40 ; le développement des techniques de microscopie électronique et d'isolement par centrifugation mettront un terme à la controverse.

Les virus d'Archéobactéries seront découverts beaucoup plus récemment puisque les premiers phages d'Archéobactérie parasitant *Halobacterium salinarium* ne seront isolés qu'au milieu des années 70 (Torvisk et Dundas 1974). Mais il faudra attendre les travaux précurseurs de Wolfram Zillig et collaborateurs, dès le début des années 80, pour que l'isolement de virus d'Archéobactéries halophiles puis hyperthermophiles prenne une véritable ampleur.

Depuis ces travaux précurseurs, ce sont près de 10000 « espèces » virales qui ont été isolées et reconnues comme valides par le comité international de taxonomie des virus (ICTV). Ces éléments ont été isolés dans presque tous les environnements. Sur le plan quantitatif, il s'agit d'un composant majeur de la biosphère puisqu'on estime par exemple que le nombre de phages infectant des bactéries dans des environnements aquatiques approche les 10 millions de particules par ml, soit au moins 10 fois plus que les organismes cellulaires (Bergh et al. 1989 ; Wommack et Colwell 2000).

II. Qu'est ce qu'un virus ?

Un virus est traditionnellement défini comme un parasite intracellulaire obligatoire, infectieux, capable de s'autorépliquer au sein de son hôte cellulaire. Un virus est composé d'une part d'un acide nucléique et d'autre part d'une ou de plusieurs protéines qui participent à la formation de la capside isolant le génome viral du milieu extérieur. Certains virus sont aussi isolés du milieu extérieur par une enveloppe contenant des lipides.

Derrière cette définition simple, se cache une extraordinaire diversité. D'abord sur la nature du génome. Les génomes viraux peuvent être circulaires ou linéaires, et être composés d'ADN simple ou double brin, ou d'ARN simple ou double brin. La taille de ce génome est extrêmement variable. Certains virus ARN simple brin (ssARN) « satellites » possèdent des génomes d'à peine quelques centaines de bases [pas plus de 0.22 kb pour le Rice yellow mottle virus (Collins et al. 1998)] tandis que certains virus « géants » double brin ADN (dsADN) possèdent des génomes de plusieurs centaines de kilobases [jusqu'à 800 kb pour le Mimivirus d'amibe (La Scola et al. 2003), plus gros que certains génomes bactériens] .

Cette diversité génomique des virus est associée à une grande diversité dans leurs stratégies réplicatives. La plupart des génomes viraux codent, à différents degrés, des éléments de leur propre machinerie de réplication et de transcription, incluant en particulier des ADN ou ARN polymérases, et détournent par contre divers composants de la machinerie de leur hôte pour compléter leur cycle. Certains virus satellites sont si simples que leurs génomes ne codent ni pour des enzymes de réplication, ni pour des protéines de structure. Ils ont donc besoin à la fois d'un hôte cellulaire et des protéines d'un autre virus co-infectant. Enfin certains génomes viraux « persistent » dans leur hôte, soit sous forme intégrée dans le génome cellulaire (prophage) soit sous forme de plasmide.

Historiquement, cette grande diversité a rendu difficile la classification de ces éléments. Jusqu'à une date récente, les critères principaux de classification retenus par l'ICTV étaient la morphologie générale et la nature de l'acide nucléique. Pourtant, le séquençage d'un nombre

croissant de génomes viraux a très largement remis en cause cette classification (Lawrence et al. 2002). Ainsi la morphologie du virion est utilisée pour grouper les virus dsADN « à queue » en trois groupes : *Siphoviridae, Myoviridae,* et *Podoviridae*. Pourtant comme on peut le voir sur la figure 13, ces familles regroupent des éléments qui ne partagent que très peu de gènes homologues, sinon aucun, entre eux (HK97 et L5). Alors que des éléments qui partagent pourtant une fraction substantielle de gènes homologues sont classés dans 2 familles différentes (P22 et HK97). De plus, les gènes codant les protéines associées à des morphologies ressemblantes sont souvent non-homologues (par exemple au sein des *Myoviridae* les phages T4 et Mu). Cela ne serait il pas le signe d'une évolution convergente, indiquant qu'il existerait des contraintes fonctionnelles favorisant une certaine morphologie de virion ?

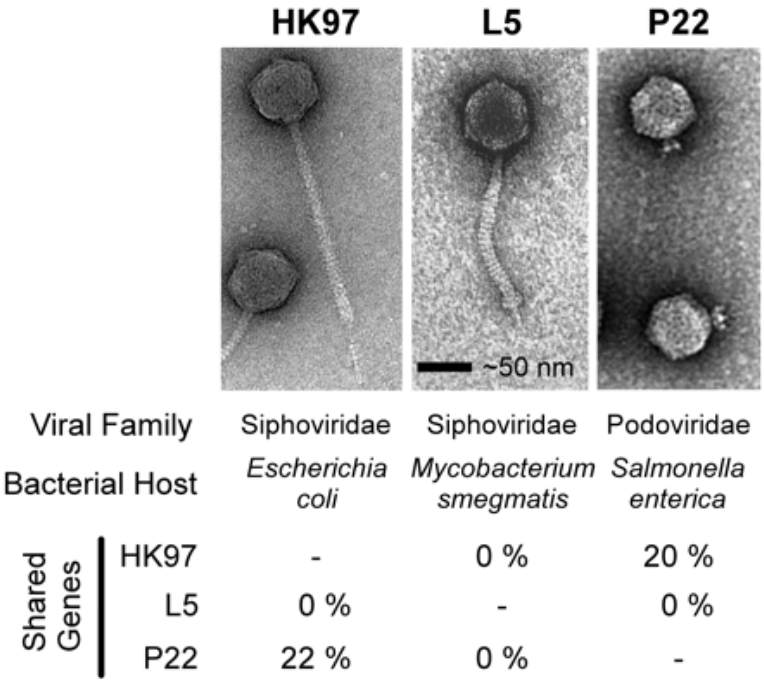

Figure 13 : Conflit entre ressemblance morphologique et ressemblance génétique. Bien que HK97 et L5 partagent une morphologie similaire, ils n'ont aucun gène homologue entre eux. Inversement, P22 a un nombre substantiel de gènes en commun avec HK97 mais ne lui ressemblent morphologiquement pas. Le critère de l'homologie est basé sur un score *E* de BLASTP (Altschul et al. 1997) meilleur que 10^{o} (d'après Lawrence et al. 2002).

III. Mécanismes d'évolution des virus.

Le séquençage d'un nombre croissant de génomes viraux a permis d'établir certains aspects clés de l'évolution des virus qui montrent que la transmission verticale de l'information est loin d'être le seul mécanisme évolutif des virus. Ces mécanismes non verticaux d'évolution sont principalement de 3 ordres :

- Le mosaïcisme dans le contenu en gènes
- La recombinaison homologue
- Les transferts horizontaux de gènes en provenance de leurs hôtes

a. Recombinaison non-homologue et mosaïcisme

Ce phénomène a été mis en évidence dès que des collections de virus suffisamment proches phylogénétiquement ont été étudiées d'un point de vue génomique. Ce phénomène a d'abord été démontré chez les phages tempérés dsADN Lambdoïdes (Highton et al. 1990). Des recombinaisons illégitimes (non homologues) entre deux génomes de virus co-infectant une cellule produisent des génomes chimériques. Ainsi lorsque l'on compare les génomes d'une collection de virus phylogénétiquement proches, on observe des réassortiments de groupes de gènes (ou modules), si bien que les génomes comparés apparaissent comme des mosaïques de modules les uns avec les autres (Figure 14).

Ce phénomène a aussi été démontré sur des collections de phages caudés tempérés de *Mycobacterium* (Pedulla et al. 2003) et de bactéries lactiques (Brüssow et Hendrix 2002). Chez des phages lytiques, dont on peut penser que les possibilités de recombinaison entre génome de phage co-infectant sont beaucoup plus faibles, le mosaïcisme a néanmoins été mis en évidence mais plus rarement, comme par exemple chez les phages apparentés à T4 (Repoila et al. 1994). Le mosaïcisme existe aussi chez les phages filamenteux ssADN (Lawrence et al. 2002). Mais il ne semble pas avoir été mis en évidence chez des Virus d'Eucaryotes ou d'Archéobactéries bien qu'a priori rien n'indique que ce phénomène ne pourrait pas être possible.

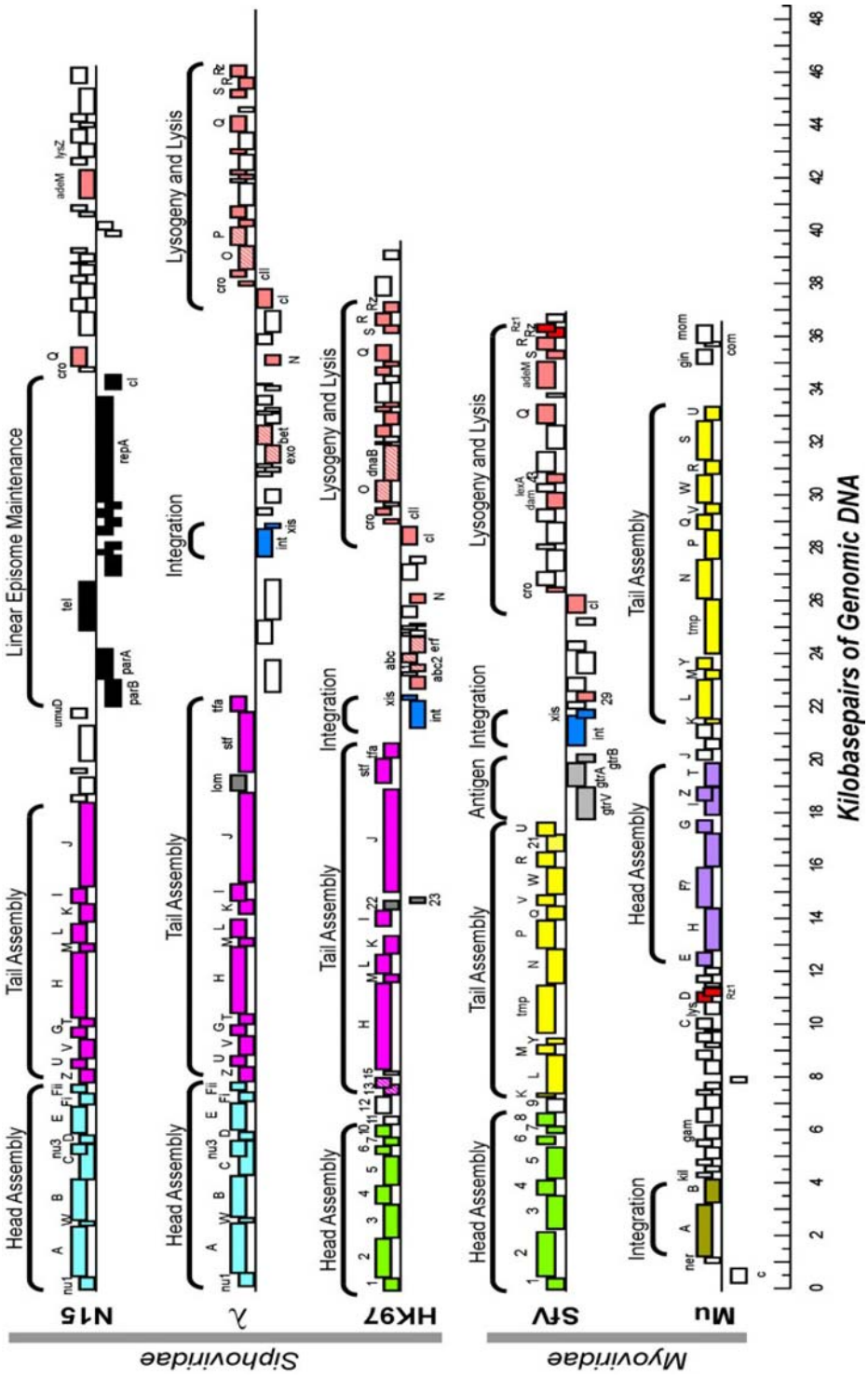

Figure 14 : Mosaïcisme en module de phage dsADN à queue infectant *E. coli*. Les modules homologues sont indiqués avec la même couleur (tiré de Lawrence et al. 2002).

b. Recombinaison homologue

Le phénomène de recombinaison entre segments homologues de deux génomes viraux (appartenant ou non à deux « espèces » virales identiques) semble être un événement fréquent puisqu'il a été mis en évidence chez la presque totalité des groupes viraux. Que ce soit des bactériophages dsADN [par exemple chez P2 (Nilsson et Haggard-ljungquist 2001)], des virus Eucaryotes dsADN [par exemple chez les Adenovirus (Nagy et al. 2002)], des virus d'Archéobactéries dsADN [Rudivirus (Peng et al. 2001)], des Rétrovirus [comme par exemple HIV (Robertson et al. 1995)] ou encore des virus ssARN d'Eucaryote (Allison et al. 2002).

Ce mécanisme pourrait permettre à certains allèles avantageux de se répandre rapidement dans une population de virus, par exemple ceux permettant d'échapper aux systèmes immunitaires de l'hôte ou d'augmenter le pouvoir infectieux (Guillot et al. 2000), ou encore d'exploiter complètement leur environnement en infectant de nouveaux hôtes.

La recombinaison homologue entre virus semble favorisée par le fait que les gènomes viraux codent souvent des enzymes de la recombinaison, ainsi le phage T4 code à la fois pour un homologue de RecA, de la protéine ssb (single strand binding protein) et d'une enzyme impliquée dans la résolution des jonctions de Holliday.

c. Transfert horizontaux de gènes

La plasticité des génomes viraux reflète aussi leur capacité à acquérir (et à perdre) des segments d'ADN de diverses origines, d'abord en provenance de leurs hôtes. Ce phénomène est historiquement connu dans le cas des virus Eucaryote oncogènes, mais le séquençage de nombreux génomes de virus dsADN de grande taille a permis de montrer qu'une large catégorie fonctionnelle de gènes pouvait être acquise par les virus via des évènements de transferts horizontaux (Baldo et McClure 1999 ; Moreira 2000 ; Hughes et al. 2002 ; Hugues et al. 2003). La figure 15 montre ainsi que les Herpesvirus ont acquis, indépendamment les uns des autres, le gène codant pour l'Interleukine IL-10 en provenance de leurs hôtes. Toutefois, si l'on compare une collection de génomes proches, on constate qu'il existe toujours un « corps » conservé de gènes dans tous les génomes viraux, et une autre catégorie regroupant des gènes qui sont présents uniquement dans quelques génomes. Ainsi, toujours chez les Herpesvirus, environ 20% des gènes présents dans un génome donné ne sont pas présents chez les autres génomes d'Herpesvirus. De plus ces gènes non-conservés codent très

souvent pour des protéines de fonction inconnue (Alba et al. 2001 ; Montague et Hutchison 2000). Cette catégorie de gènes peu conservés est souvent interprétée comme issue de transferts horizontaux récent de gènes en provenance de l'hôte. Pourtant cette explication apparaît un peu simpliste, car souvent ces gènes n'ont pas d'homologue dans le génome des hôtes, ou bien cet homologue est phylogénétiquement très éloigné. Cette situation, fréquente chez les virus dsADN possédant un gros génome, pose la question de l'origine évolutive de ces gènes.

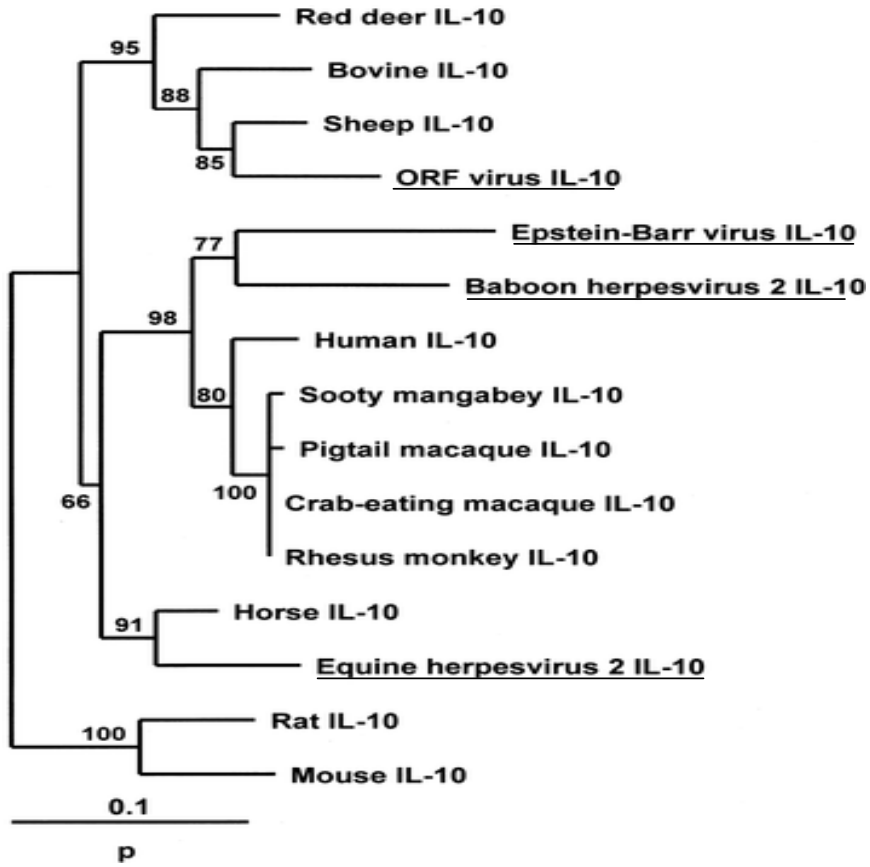

Figure 15 : Arbre raciné de distance (NJ) de l'Interleukine 10 chez les mammifères et les Herpesvirus (soulignés). Les Herpesvirus sont largement polyphylétiques et se positionnent dans chaque cas comme groupe frère de leurs hôtes respectifs. Cette phylogénie supporte l'idée de multiples transferts horizontaux récents du gène, polarisés de la cellule vers le virus (issu de Hughes et al. 2002).

Chez les Bactériophages, la comparaison des génomes des phages Lambdoïdes a montré que dans les "clusters" de gènes codant pour les protéines de structures, on observait là aussi des gènes peu conservés, présents d'une manière erratique dans tel ou tel génome (Hendrix et al. 2000). Ces gènes étant très souvent de fonction inconnue et d'origine évolutive mystérieuse. Cette observation a amené une théorie, la théorie des « morons » pour « more DNA ». Cette hypothèse postule que les bactériophages évoluent par des étapes de complexification croissante par insertion de fragments d'ADN (Hendrix et al. 2000). Cette hypothèse est en adéquation avec une théorie plus large qui voudrait que les virus soient des fragments d'ADN d'origine cellulaire qui auraient acquis leur autonomie et auraient évolué par étapes de complexification progressive par accrétion de gènes d'origine cellulaire (Iyer et al. 2001 ; Hendrix et al. 2000). Cette théorie, très « populaire », repose toutefois sur des arguments assez minces puisque comme nous le verrons par la suite (i) la présence d'homologues cellulaires de gènes viraux n'impliquent pas que les virus ont acquis ces gènes en provenance de leurs hôtes et non l'inverse (ii) une grande partie des gènes viraux n'ont pas d'homologues cellulaires (iii) les phénomènes observés peuvent s'expliquer d'une manière aussi parcimonieuse par des pertes de gènes et des simplifications à partir d'un ancêtre suffisamment complexe. De fait, il semble peu contestable qu'une fraction des gènes présents dans les génomes viraux n'a pas été acquise récemment par les virus en provenance de leurs hôtes. Ces gènes témoigneraient ils d'une origine « ancienne » des virus ?

IV. Origine(s) et histoires évolutives des virus

a. Les virus sont probablement polyphylétiques

Etant donné la très grande diversité des génomes des virus, il semble très peu probable que tous ces éléments aient une origine unique (Iyer et al. 2001). Au contraire, un consensus se dégage autour d'une origine multiple des virus. Toutefois, des analyses comparées de génomes de virus ont parfois suggéré l'existence d'une origine commune entre des virus très divergents. Ainsi certains virus dsARN et dsADN de petite taille codent pour des enzymes de la réplication qui sont homologues, en particulier des Hélicases (Gorbalenya et al. 1990). De même, certains virus ssADN et des plasmides dsADN utilisent une protéine homologue pour l'initiation de la réplication par le mécanisme du « cercle roulant » (Koonin et Ilyina 1992).

b. Des liens évolutifs entre virus, plasmides, transposons...

Des connections évolutives ont souvent été mises en évidence entre les différents éléments génétiques mobiles des génomes et des virus. Par exemple, l'organisation génomique et les phylogénies des principaux gènes conservés chez les rétrovirus et les rétrotransposons indiquent une origine commune (Xiong et Eickbush 19990 ; McClure 1991 ; Capy et al. 1996 ; Lerat et Capy 1999). Beaucoup de virus passent dans leur cycle biologique par une forme plasmidique, et il est possible d'établir de nombreuses connections évolutives entre plasmides et virus. Ainsi, certains virus, comme les Adenovirus, et certains bactériophages, comme phi-29, partagent avec les plasmides linéaires de mitochondrie une ADN polymérase particulière que l'on ne retrouve dans aucun autre génome (Knopf 1998). Il existe aussi des éléments chimériques : ainsi le bactériophage N15, qui possède un génome composé à 50% de gènes de type bactériophage et à 50% de gènes dérivant de plasmides linéaires (voir figure 11)(Ravin et al. 2000). Enfin certains éléments, comme R391, possèdent a la fois des caractéristiques de phage tempéré (comme le mécanisme et les enzymes d'intégration du génome), des caractéristiques de plasmide (comme le mécanisme et les enzymes de transfert conjugatif de gènes) et même de nombreux transposons intégrés dans le même génome (Boltner et al. 2002).

Ainsi un rétrovirus perdant sa capacité infectieuse pourrait devenir un rétrotransposon, ou un plasmide se recombinant avec un phage pourrait donner un nouveau virus capable d'infecter d'autres cellules. Ces processus pouvant être dans les deux directions possibles.

c. Les virus sont ils issus d'ADN cellulaire ?

L'hypothèse la plus largement acceptée pour expliquer l'origine des virus propose qu'ils sont issus d'un fragment d'ADN d'origine cellulaire s'étant échappé, devenu autonome et infectieux. Proposée initialement par Howard Temin durant les années 70, cette hypothèse a souvent été reprise, l'argument principal étant que l'on trouve beaucoup de gènes homologues entre virus et cellule. Or un tel phénomène d'autonomisation d'ADN cellulaire n'a pour l'instant jamais été démontré expérimentalement, il s'agirait donc d'un événement très rare (Villarreal 1999). L'hypothèse alternative postule que les virus actuels, ou certains d'entre eux résulteraient de l'extrême réduction d'un génome de cellule ou de proto-cellule (Bandea 1983). Toutes ces hypothèses échouent néanmoins a expliquer l'origine évolutive d'un grand nombre de gènes n'ayant aucun homologue cellulaire (qu'ils codent ou non pour des protéines de fonction connue). L'existence de tels gènes accrédite plutôt l'idée de l'existence d'un « monde des virus », différentié d'un point de vue originel du « monde des cellules ». Toutefois il semble réaliste d'imaginer que les génomes des virus sont des éléments chimériques, issus à la fois de gènes d'origine cellulaire (incluant par exemple des gènes issus de transferts horizontaux en provenance des cellules, voir chapitre III partie c.) et de gènes à proprement parler « viraux » (incluant par exemple les gènes sans homologue cellulaire).

d. Les virus comme éléments « anciens » :

Un consensus semble émerger aujourd'hui sur l'existence de connections évolutives entre virus infectant les 3 domaines du vivants. En effet, des études génomiques, et surtout structurales, apportent des arguments très solides sur une origine ancienne des virus, probablement antérieure à la divergence des trois domaines du vivant. Ainsi, la structure tridimensionnelle de la protéine de capside P3 des Adenovirus (virus eucaryotes) et Hexon du phage PRD1 (phage de Bactérie) montre de très fortes ressemblances bien qu'aucune similarité au niveau de la séquence primaire ne soit détectable (Benson et al. 1999)(figure 16). Ces gènes ne possèdent pas d'homologue cellulaire connu à ce jour, ce qui rend peu probable leur acquisition indépendante par transferts horizontaux en provenance de leur hôte. Il est donc fort possible que ces structures soit homologues et héritées d'un ancêtre commun

antérieurement à la divergence des Bactéries et des Eucaryotes. De plus ces deux virus partagent une ADN polymérase particulière (voir chapitre C) et une organisation génomique similaire (Bamford et al. 2002). Néanmoins il est possible, dans ce cas, d'imaginer des convergences évolutives et/ou des acquisitions indépendantes de gènes cellulaires.

D'autres similarités structurales ont aussi été mises en évidence entre des virus dsARN du groupe des Réovirus (Virus Eucaryotes) et le phage phi6 (phage de Bactérie), entre les phages à queue des 3 domaines (voir pour revue Bamford et al. 2002) ou encore entre le bactériophage ssADN phiX174 et certains virus ssARN de plantes et de champignons. Enfin, de très nombreuses protéines de structure de virus infectant les 3 domaines partagent un motif particulier appelé « Jellyroll » que l'on ne retrouve chez aucune protéine cellulaire (figure 16B), ce qui accrédite l'idée que les virus sont au moins en partie constitués d'acide nucléique ancien, différencié de l'ADN cellulaire [ce que Dennis Bamford nomme le « self » des virus (Bamford et al. 2002)] et ayant divergé sans doute avant le LUCA.

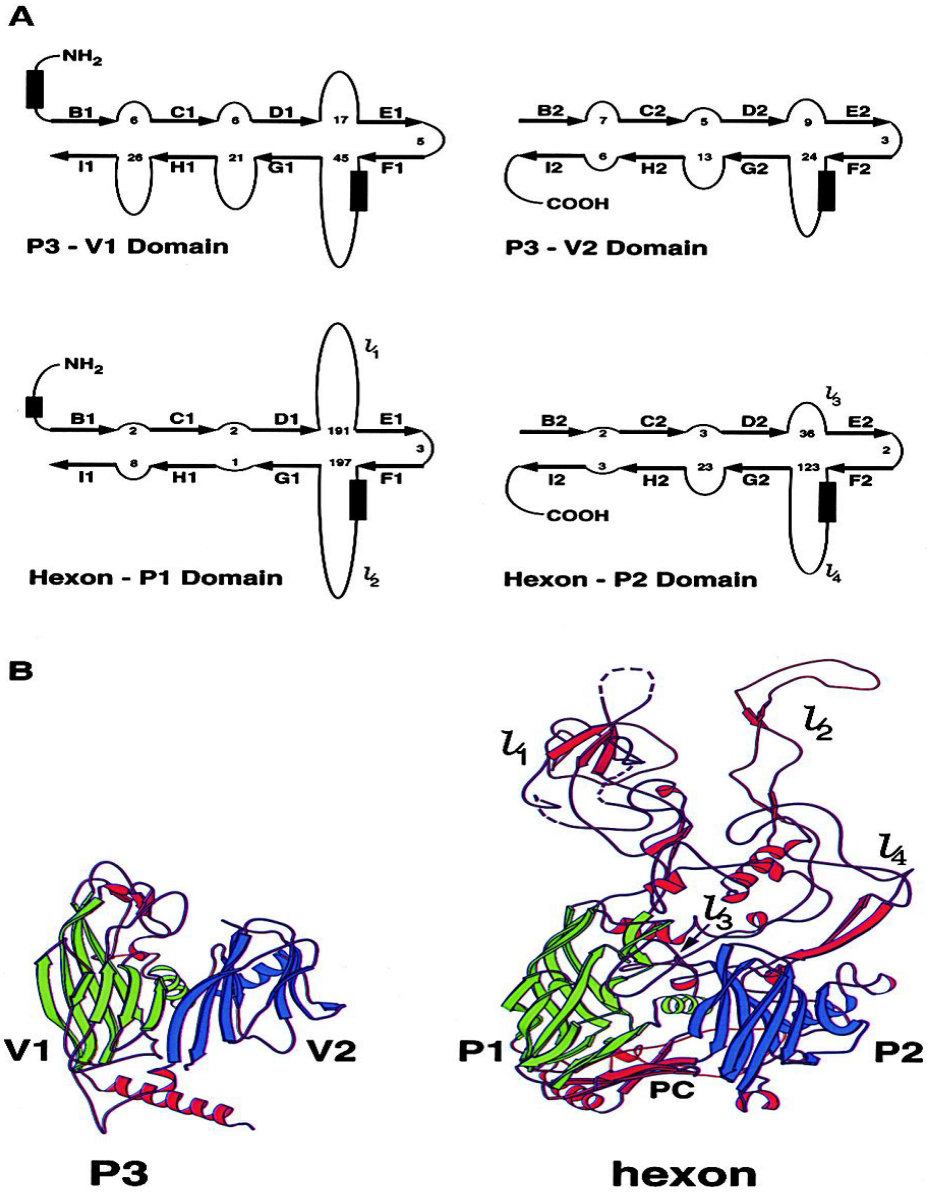

Figure 16 : Comparaison des monomères de la protéine P3 du bactériophage PRD1 et de la protéine hexon chez un Adenovirus. A : les flèches et les rectangles noirs indique respectivement les feuillets β et les hélices α. B: Les motifs Jellyroll de chaque monomère sont respectivement noté V1/V2 et P1/P2. D'après Benson et al. 1999.

Au niveau de l'architecture génomique de fortes ressemblances existent aussi entre les virus d'Archéobactéries SIRF et SIRV1 et des virus eukaryotes (Prangishvilli et al. 2001)(Blum et al. 2001). Mais à ce niveau, il est souvent très difficile de trouver des caractères homologues dès que l'échelle évolutive est trop grande. Au sein des virus dsADN d'Eucaryotes et de Bactéries plusieurs tentatives de comparaison de génomes pour différents groupes de virus ont été publiées. Chez les virus Eucaryotes, sur la base du contenu en gènes de leurs génomes Iyer et collègues ont proposé que 4 groupes de virus, Poxvirus, Asfarvirus (parasitant tous les deux des vertébrés), Iridovirus (Virus de Poissons et d'Insectes) et Phycodnavirus (virus d'Algues) partageaient une origine commune (Iyer et al. 2001). Un scénario évolutif a été proposé pour rendre compte de l'évolution des génomes de ces virus (figure 17).

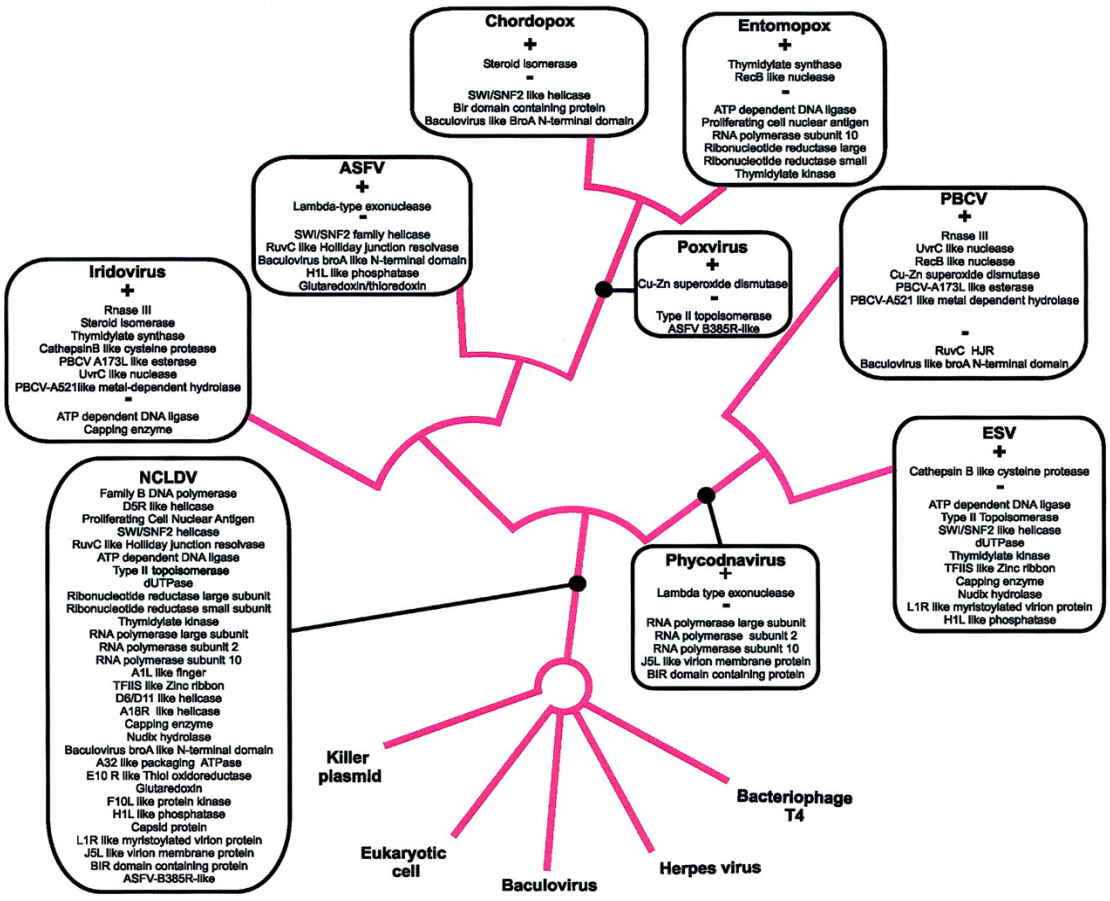

Figure 17 : Cladogramme des Virus dsADN Eucaryotes. L'ancêtre commun est annoté comme « NCDLV ». Le contenu en gènes des génomes est indiqué dans les cercles, « + » signifie un gain des gènes dans la lignée, « - » indique une perte. Tiré de Iyer et al. 2001.

Toutefois ce scénario suppose que la présence d'un gène conservé chez ces 4 groupes de virus résulte d'un héritage vertical de leur ancêtre commun. Or, pour beaucoup de ces gènes qui ont des homologues cellulaires chez leurs hôtes, l'acquisition multiple et indépendante par transferts horizontaux ne peut pas être rejetée. Seule une analyse phylogénétique détaillée de chacun des gènes pourrait apporter une réponse. Néanmoins ce scénario indique que, si l'on possède une collection suffisamment détaillée de génomes de virus, malgré leur plasticité (voir partie III), il est possible de proposer des scénarios retraçant pas à pas leur évolution.

Pour les bactériophages, un scénario évolutif à grande échelle a été proposé sur la base de la comparaison de 105 génomes complets (Rohwer et Edwards 2002). Aucun gène n'est conservé au sein de ces 105 génomes : une approche par compatibilité a donc été proposée en posant l'hypothèse que, plus deux organismes partagent de gènes homologues en commun, plus ils sont proches d'un point de vue évolutif : il en résulte la figure présentée en figure 18. Cette technique se révèle efficace pour grouper des phages ayant des génomes très proches alors que leurs hôtes sont éloignés phylogénétiquement (ainsi par exemple le phage SIO1 qui infecte une alpha-protéobactérie marine, est regroupé avec le phage T7 qui infecte des gamma-protéobactéries du tube digestif humain). Mais cette technique n'est pas robuste à grande échelle évolutive car à ce niveau la conservation des gènes entre les groupes de virus est très faible, et l'impact de transferts horizontaux de gènes indépendants entre les lignées affecte très fortement la topologie de l'arbre (Rohwer et Edwards 2002). Cette approche non cladiste au sens de Hennig (voir chapitre A, partie I), pose explicitement le problème de la reconstruction de l'histoire évolutive des virus avec les outils traditionnellement utilisés pour étudier l'évolution des êtres cellulaires. Les outils sont ils bien adaptés ? Quelle place pour le transfert vertical de l'information en comparaison du transfert horizontal ?

Pris ensemble, tous ces résultats indiquent que des virus très divergents, infectant des hôtes phylogénétiquement très éloignés, partagent des caractères homologues. Même si l'on ne peut pas exclure des phénomènes de convergence évolutive, il est très probable qu'au moins une partie de ces caractères sont bien hérités d'un ancêtre commun. Ils sont donc au moins en partie constitués d'éléments très anciens, qui trouvent probablement leur origine antérieurement à la divergence des trois domaines. Cette origine ancienne permet en outre d'aborder différemment les deux visions antagonistes sur l'origine des virus : celle des tenants d'un monde des virus différentié du monde des cellules [illustré avec le concept du "self" de Dennis Bamford (Bamford et al. 2002)] et celles des tenants d'une origine cellulaire des virus

soit par régression parasitaire (Bandea 1983) soit par autonomisation d'un morceau d'ADN et accrétion successive de "morons" (Hendrix 2000). En effet on peut penser que les virus sont originaires de cellules antérieures au LUCA (Forterre 1992), soit par autonomisation soit par régression, et pourquoi pas issus de lignées cellulaires ayant disparu. Les génomes des virus actuels auraient conservé des gènes d'origine cellulaire qui furent perdus dans les lignées cellulaires. Inversement, les gènes viraux ayant des homologues cellulaires seraient ou bien des gènes très anciens, conservés dans les lignées cellulaires et virales, ou bien acquis par les virus par un processus d'accrétion de gène cellulaire. Néanmoins la très longue histoire évolutive des virus pourrait aussi les avoir amenés à inventer des gènes, des fonctions ou des structures sans homologue cellulaire. Malheureusement l'absence de caractère homologue conservé chez tous les virus, et notre connaissance trop partielle de leur biodiversité (Breitbart et al. 2002) rend très difficile la reconstitution de l'histoire évolutive des virus à une échelle de temps très ancienne. Il est possible que le séquençage d'un beaucoup plus grand nombre de génomes viraux nous permettra de reconstruire pas à pas cette évolution en reliant entre eux des virus de plus en plus divergents. Il est aussi possible qu'en se focalisant sur certains gènes particuliers, les plus conservés, ou en reconnaissant des gènes homologues présent uniquement chez certains virus, on puisse reconnaître des synapomorphies qui nous permettront de regrouper certains virus. Toutefois, il semble aussi très important de mieux comprendre les relations évolutives entre les virus et leurs hôtes cellulaires. En particulier les flux de gènes qui semblent exister entre ces entités et qui participent à rendre très ardue la reconstruction de leurs histoires évolutives.

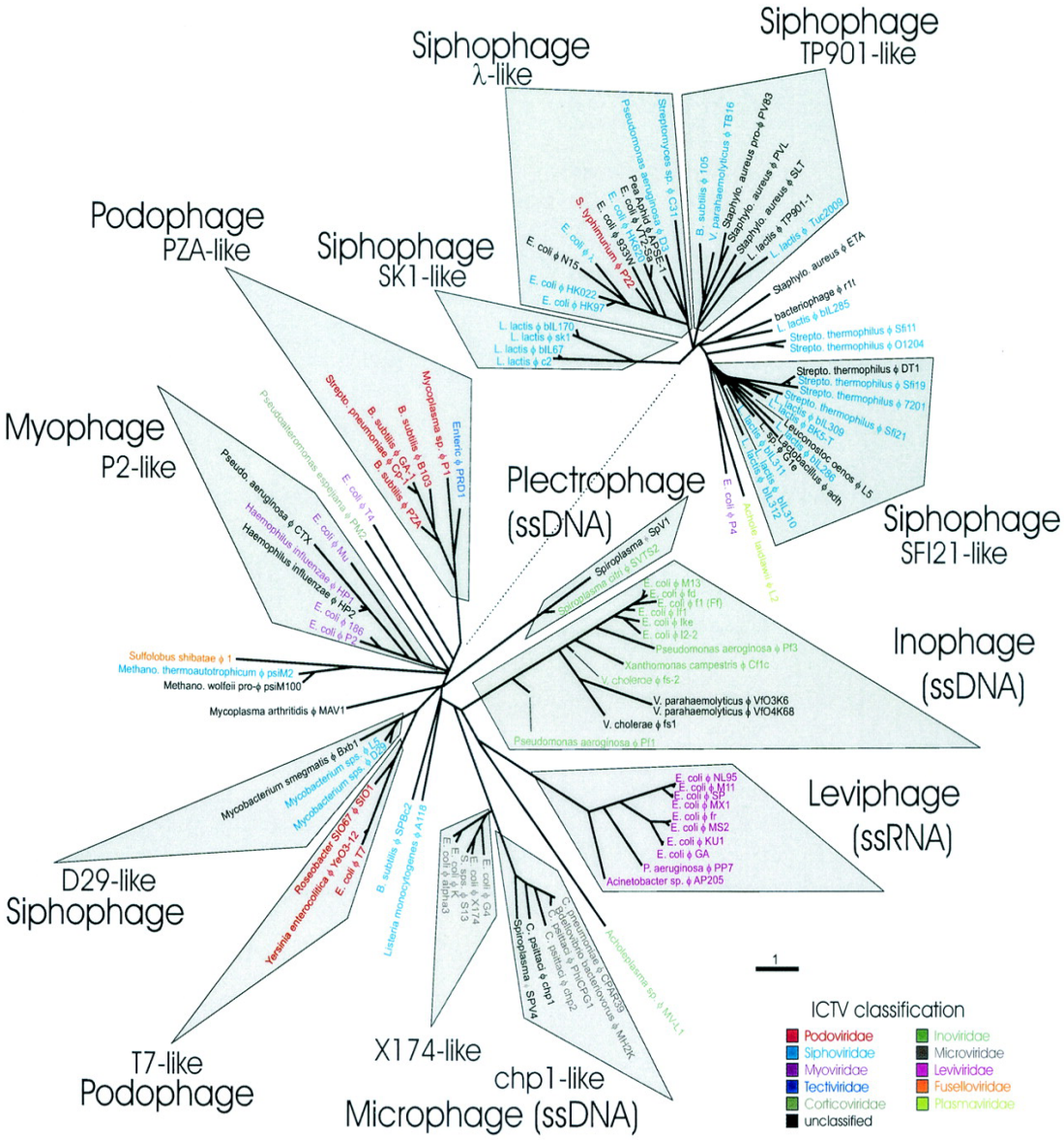

Figure 18 : Arbre protéomique des phages issu de la comparaison de 105 génomes. Chaque phage est colorié en fonction de sa classification selon ICTV. Pour plus de clarté le plus grand groupes, celui des Siphophages a été détaché de l'arbre pour le rendre plus lisible, la branche en pointillé montre son branchement effectif (d'après Rohwer et Edwards 2002).

V. La relation Hôte/Parasite entre virus et cellule

Certains virus, et singulièrement les virus à ADN, tendent à établir des relations à long terme avec leur hôte, sans provoquer leur mort, mais plutôt en persistant au sein de ces dernier par exemple en intégrant leur génome dans le génome hôte. D'autres, au contraire, se répliquent très vite au sein de leur hôte après l'avoir infecté et sont transmis à d'autres cellules à très court terme. Ces stratégies induisent, de part et d'autre, l'existence de mécanismes moléculaires spécifiques et complexes, d'interactions entre les virus et leurs hôtes, dont les principaux seront présentés dans ce chapitre.

a. Les mécanismes moléculaires de la persistance des virus

Les mécanismes moléculaires de la persistances chez les virus ont été précisément étudiés chez les Bactériophage lactiques où l'on disposait de génomes très proches, dont les uns appartenaient à des éléments lytiques (donc non persistants) et les autres à des éléments lysogènes, pouvant persister en intégrant leur génome dans celui de leurs hôtes (voir pour revue Brussow et al. 2001). Il est rapidement apparu que ces phages lytiques résultaient de la perte de modules entiers (cassettes de gènes) ou de la perte d'un ou de plusieurs gènes particuliers par délétion. Dans le module on trouve le plus souvent une intégrase qui permet d'intégrer le génome phagique d'une manière site-spécifique par recombinaison (le plus souvent au niveau d'un ARN de transfert), une excisase qui permet au génome viral de "sortir" du génome cellulaire, des répresseurs spécifiques de la transcription des gènes phagiques, et parfois un ou plusieurs gènes codant pour des fonctions de défense immunitaire protégeant la bactérie hôte contre une sur-infection par des phages lytiques (le plus souvent des endonucléases de restriction site-spécifiques, qui sont utilisées pour dégrader l'ADN d'un autre phage co-infectant la cellule hôte).

D'autres phages tempérés utilisent des mécanismes d'intégration aléatoire, catalysés par des transposases ou se maintiennent sous forme de plasmides circulaires autoréplicatifs.

L'intégration d'un prophage dans un génome cellulaire semble être un phénomène très fréquent puisque tous les génomes de Bactéries séquencés à ce jour contiennent au moins un prophage intégré.

b. Les mécanismes moléculaires des stratégies lytiques

Les virus lytiques codent souvent pour plusieurs enzymes impliquées dans la réplication et le métabolisme de l'ADN. Ils ont effet développé des systèmes moléculaires pour amplifier la réplication de leur propre ADN (Villareal 1999) :

- Soit en inventant de nouveaux systèmes de réplication capables de reconnaître et de dupliquer spécifiquement leur ADN : par exemple l'ADN polymérase des Adenovirus appartenant à la famille B, qui utilise un mécanisme d'initiation de la réplication différent de celui des cellules hôtes (Knopf, 1998), ou encore la réplication en « cercle roulant » de nombreux plasmides et virus à génome circulaire (Koonin et Ilyina 1992).
- Soit en dégradant l'ADN de leur hôte avec des endonucléases tout en protégeant leur propre ADN via des modifications de celui-ci (par exemple l'Hydroxyméthylcytosine qui remplace la cytosine dans le génome des phages T-pairs) ou en isolant leur ADN avec des protéines qui se fixe spécifiquement sur leur génome (cas des Herpesvirus).
- Soit en détournant l'appareil de réplication de l'hôte via, par exemple, l'invention de protéines qui se fixent spécifiquement à l'origine de réplication du virus et qui ont une très forte affinité avec les composants de la réplication de l'hôte (cas de l'antigène T chez le virus SV40 qui lie spécifiquement la primase Eucaryote, ou encore de la protéine P du bactériophage λ qui recrute spécifiquement l'Hélicase Réplicative DnaB).

D'une manière non spécifique les virus peuvent aussi augmenter la synthèse globale d'ADN de leur hôte en dérégulant le cycle cellulaire. C'est le cas par exemple de certains virus oncogènes qui peuvent induire la prolifération des cellules les hébergeant (Adenovirus, Hepadnavirus, Herpesvirus…).

Enfin beaucoup de génomes de phages lytiques codent pour des endonucléases de restriction site-spécifiques qui sont aussi utilisées pour dégrader l'ADN d'autres phages co-infectant la cellule hôte. Certains virus Eucaryote, les Phycodnavirus, possèdent aussi des endonucléases spécifiques utilisées pour dégrader à la fois le génome de l'hôte et celui d'un virus co-infectant. Ces virus protègent leur propre ADN avec des ADN méthyltransférases qui rendent ineffective l'action des endonucléases sur les sites dont les bases adénine ou cytosine sont méthylées (Van Etten et Meint 1999) . Il est à noter que ces endonucléases n'ont d'homologue cellulaire connu que chez les Bactéries et non pas chez les Eucaryotes comme on pourrait s'y attendre pour un virus Eucaryote.

c. Les réponses immunitaires des procaryotes

Chez les procaryotes, la réponse immunitaire contre les phages consiste essentiellement en des systèmes de restriction et de modification (RM) de l'ADN (Sowers 1995). Il s'agit d'un système qui comprend d'une part une endonucléase qui clive l'ADN d'une manière site spécifique et d'autre part une enzyme qui modifie certaines bases de l'ADN au niveau de ce site pour les protéger de la coupure par l'endonucléase qui lui est associée dans le système. Il existe un grand nombre de systèmes RM différents, non-homologues dont l'origine évolutive est mal connue (Roberts et Macelis 1997). Toutefois, nous avons vu précédemment que beaucoup de virus possédaient leur propre système RM. Il est possible que les cellules aient « recruté » ces systèmes viraux via des événement de transfert horizontau. Cette hypothèse est confortée par le fait que beaucoup de systèmes RM sont portés soit par des plasmides (voir plus bas), soit par des élément ressemblant à des prophages cryptiques (Anton et al. 1994) soit encore des transposons (Brassard et al. 1995).

Beaucoup de Bactéries possèdent des plasmides portant un système RM. C'est le cas d'*E. coli* et son système *Eco*RI. Des expériences ont montré que le remplacement de ce système par un autre système RM plasmidique était très difficile. La transformation *in vitro* est d'une très faible efficacité et la plupart des colonies survivantes ont une croissance très ralentie (Naito et al. 1995). L'explication proposée est qu'après le remplacement du système « A » par le système « B » il ne reste plus assez d'enzyme de modification « A » pour protéger tous les sites de l'ADN néo-synthétisé vis à vis du pool restant d'enzyme de restriction « A ». Ce qui entraîne de nombreuses coupures de l'ADN et la mort des cellules. Ainsi, les virus persistants et plasmides utiliseraient ces système RM pour se maintenir d'une manière très efficace dans les génomes en tant qu'élément égoïstes pour deux raisons (Figure 19) :

- détruire l'ADN de phages et plasmides co-infectants dès le début de leur infection
- détruire son hôte si la co-infection aboutit a l'élimination de l'élément de départ (d'un point de vue populationel, le virus se maintenant uniquement dans les hôtes qui ne sont pas co-infectés)

Toutefois il peut aussi être avantageux pour l'hôte de conserver un plasmide ou un prophage intégré portant un système SM pour éviter les surinfections et les effets délétères de la colonisation du génome par ces éléments (Kobayashi 1998).

Néanmoins il existe aussi chez les bactéries des enzymes de modification en « solo » qui pourrait être utilisées préventivement pour se protéger des enzymes de restriction (Kobayashi 1998). On a aussi mis en évidence chez *E. coli* des endonucléases méthyl-spécifiques en «solo» qui clivent sélectivement des bases méthylées (Raleigh et Wilson 1986). En réponse à l'infection par des éléments portant un système RM, cette enzyme pourrait constituer un système de défense supplémentaire.

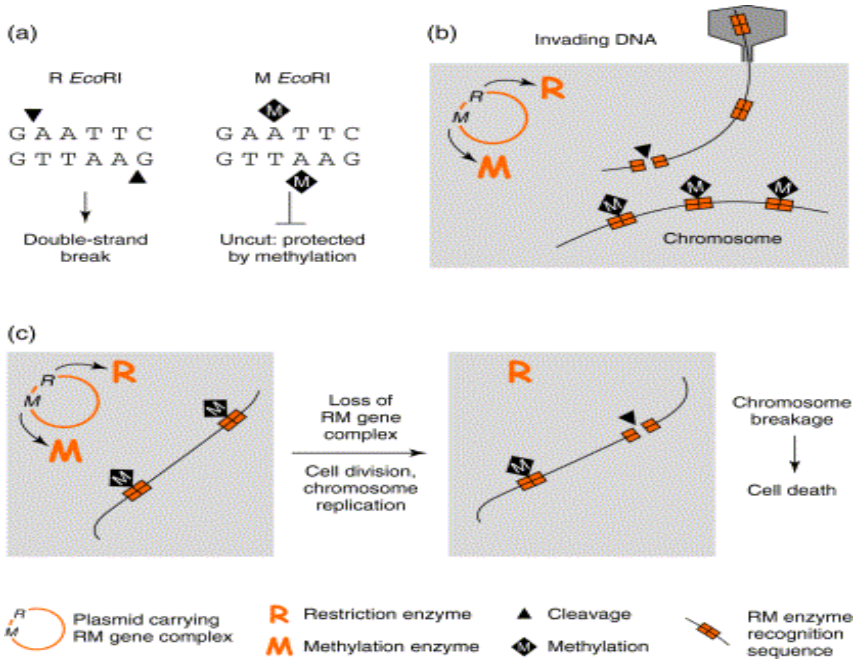

Figure 19 : Le système de restriction/modification des Procaryotes et son mécanisme de persistance en tant qu'élément « égoïste » dans les génomes. (a) Enzymes et réactions, les sites de clivage et de méthylation sont indiqués. (b) Attaque par un bactériophage à ADN non modifié mais avec chromosome modifié. (c) Perte du système RM initial et mort de la cellule du fait de la subsistance d'un pool d'endonucléase de restriction. (Tiré de Kobayashi 1998)

d. La réponse immunitaire chez les Eucaryotes

Les Eucaryotes n'utilisent pas de systèmes RM, ils ne possèdent pas d'enzyme de restriction, par contre leurs génomes contiennent par contre des enzymes de méthylation des acides nucléiques qui sont utilisée pour différents processus, dont la régulation globale de l'expression des gènes. D'un point de vue évolutif, il est possible que ces enzymes aient pu

servir initialement pour se protéger des enzymes de restriction des virus (Doerfler 1996) ou pour empêcher la colonisation du génome par des éléments transposables en réprimant globalement l'expression des gènes, y compris ceux des transposons (McDonald 1998 ; Bowen et Jordan 2002) .

Face à une infection virale, la réponse immunitaire la plus courante chez les Eucaryotes est l'apoptose, c'est à dire le suicide de la cellule infectée. Beaucoup d'Eucaryotes présentent des cellules différenciées qui ne se divisent plus et ne répliquent plus leur ADN. Lorsque ces cellules sont infectées par un virus ce dernier induit la mobilisation de la machinerie de réplication de l'ADN pour répliquer son propre ADN, ceci a pour effet d'activer le processus d'apoptose par stabilisation ou accumulation d'un composant clé : le facteur de transcription p53 (Toeodoro et Branton 1997). Beaucoup de virus bloquent l'apoptose très en amont en produisant des protéines qui fixent p53 et l'inactivent. Ainsi par exemple l'antigène T du virus SV40 se lie spécifiquement à p53 et l'empêche de se fixer aux séquences promotrices des gènes impliqués dans les premières étapes de l'apoptose (Bargonetti et al. 1992). D'autres stratégies sont utilisées par les virus pour bloquer plus en aval l'apoptose, voir Toeodoro et Branton (1997) pour revue exhaustive de ces mécanismes.

Néanmoins, là aussi, la relation entre le virus et son hôte peut être plus complexe. Ainsi, au cours des étapes finales de son cycle, il peut être avantageux pour un virus, non plus de bloquer mais d'induire l'apoptose. En effet l'apoptose est utilisée par les virus pour provoquer la mort de la cellule, la dégradation des membranes et la libération des virus nouvellement produis. Beaucoup de génomes de virus codent ainsi pour des protéines connues pour induire l'apoptose, souvent par des voies indépendantes de p53. Pour revue de ces protéines voir Toeodoro et Branton (1997).

Les vertébrés ont aussi développé un système immunitaire spécifique pour combattre les agents pathogènes, dont les virus. Il sort du cadre de ce manuscrit de décrire les différentes étapes de la réponse immunitaire mais il peut être intéressant de dire quelques mots à propos de l'origine évolutive de certains composants de cette machinerie cellulaire. La réponse immunitaire spécifique chez les vertébrés passe par la production d'immunoglobulines. Les immunoglobulines sont codées par des loci complexes composés de segments discrets qui subissent des recombinaisons pour donner un énorme répertoire de combinaisons possibles. Ces recombinaisons sont catalysées par deux enzymes RAG1 et RAG2. Le mécanisme de recombinaison est très similaire de celui de la transposition des transposons de classe II et de l'intégration site-spécifique de bactériophages tempérés comme λ (Lewis et Wu 1997). De

plus il a été démontré que RAG1 se fixe sur une séquence très proche des séquences cibles des intégrases site-spécifiques (Simon et al. 1980), et que si l'on remplace le domaine de fixation à l'ADN de RAG1 par le domaine de fixation à l'ADN d'une intégrase site-spécifique, l'enzyme RAG1 demeure complètement fonctionnelle (Spanopoulou et al. 1996). Enfin RAG1 et RAG2 catalysent effectivement ensemble la transposition in vitro (Agrawal et al. 1998). Ces résultats suggèrent que toutes ces enzymes partagent peut être une origine évolutive commune et on pourrait imaginer que les vertébrés auraient recruté ces fonctions en provenance d'un élément génétique mobile (transposon ou virus) (Spanopoulou et al. 1996 ; Bowen et Jordan 2002).

Tous ces résultats démontrent que la relation hôte/parasite entre virus et cellule ne peut pas se résumer seulement à « une course aux armements ». Dans plusieurs cas, nous avons vu que les cellules pouvaient potentiellement tirer parti des virus pour accomplir des fonctions cellulaires. Dans le chapitre V de la partie A, nous avons vu que les virus pouvaient constituer de très bons vecteurs pour les transferts horizontaux de gènes entre espèces. Ceci amène naturellement à la question suivante : d'un point de vue génétique, quel pourrait avoir été le rôle joué par les virus sur l'évolution de leurs hôtes ?

VI. La contribution des virus à l'évolution de leurs hôtes

Le séquençage d'un nombre important de génomes de Bactéries a mis en évidence l'importance des transferts horizontaux de gènes au cours de l'évolution de ces organismes (Ochman et al. 2000). Dans de nombreux cas, les gènes transférés ont été identifiés comme issus d'éléments génétiques mobiles (virus, plasmides, transposons). Souvent ces gènes ont un fort impact sur le phénotype de la souche bactérienne : pathogénicité, résistance aux antibiotiques ou propriété métabolique particulière. Ainsi de nombreux gènes impliqués dans la virulence d'une souche sont regroupés en "îlots de pathogénicité" localisés au sein de séquences codant pour des ARNt. Or de nombreux phages tempérés sont connus pour s'intégrer au sein des séquences d'ARNt. Au moins un cas est parfaitement documenté chez *Staphylococcus aureus,* où un îlot de pathogénicité est porté par un bactériophage (Lindasy et al. 1998). Parfois, la virulence est associée à un seul gène issu d'un bactériophage tempéré : par exemple, le gène codant pour l'exotoxine A chez *Streptococcus pyogenes* porté par le bactériophage T12 (Weeks et Ferretti 1984), les gènes codant pour les Shiga-toxines chez certaines souches d'*E. coli* (Jackson et al. 1987) ou encore la toxine cholérique chez *Vibrio cholerae* codée par le génome du phage phi-CTX (Waldor et Mekalanos 1996). Les gènes de résistance aux antibiotiques sont presque systématiquement codés par des éléments génétiques mobiles, plasmides mais aussi transposons comme par exemple l'élément Tn5 qui confère une grande variété de Bactéries une résistance à la fois à la kanamycine, bléomycine et streptomycine (Berg 1989). Enfin parfois les transferts de gènes issus d'éléments génétiques mobiles impliquent des grandes régions codant plusieurs dizaines de gènes tels les "îlots de symbiose". Ainsi chez *Mesorhizobium loti*, un îlot de symbiose de plus de 500kb, codant pour près de 50 protéines impliquées dans la fixation de l'azote atmosphérique et la nodulation, est intégré dans un ARNt et présente un usage des codons et des taux de G+C atypiques (Kaneko et al. 2000).

D'un point de vue quantitatif enfin, il semble que, chez certaines Bactéries au moins, une fraction très importante des gènes identifiés comme issus de transferts proviennent d'éléments mobiles (Figure 20). C'est par exemple le cas chez *E. coli*, *Synechocystis* PCC6803, *Helicobacter pylori*, ou *Archeoglobus fulgidus* (Ochman et al. 2000).

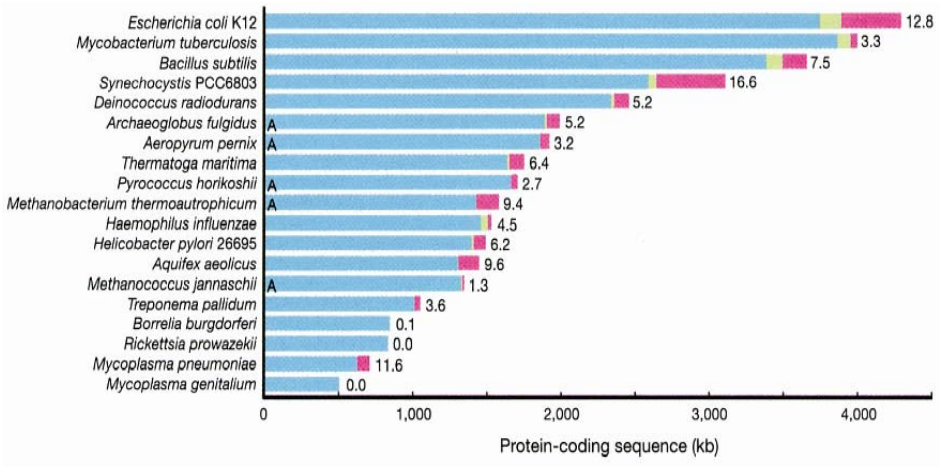

Figure 20 : Représentation graphique de la quantité de gènes "natifs" d'un génome (en bleu), de gènes détectés comme issus de transferts récents soit en provenance de phages et autres éléments génétiques mobiles (en jaune) ou d'origine inconnue (en rouge). Le pourcentage de gènes d'origine "étrangère" est indiqué par un chiffre en pourcentage du génome total. "A" indique les Archéobactéries. Extrait de Ochman et al. 2000.

Il faut toutefois noter qu'un grand nombre de gènes identifiés comme issus de transferts horizontaux ne sont pas assignables à une fonction particulière et ne sont pas intégrés dans un génome d'élément génétique mobile. Ces gènes sont souvent identifiés du fait d'un taux de G+C différent de celui du génome (voir chapitre V partie b.). Très souvent ces gènes possèdent un taux de G+C inférieur à celui de leurs hôtes, surtout en troisième position, mais aussi, d'une manière moins significative, en première et en deuxième position (figure 21). Ce résultat est vrai aussi pour des Bactéries dont les génomes ont des taux de G+C relativement bas comme *Streptococcus* et *Helicobacter* (Medigue et al. 1991 ; Lawrence et Ochman 1997, Daubin 2003).

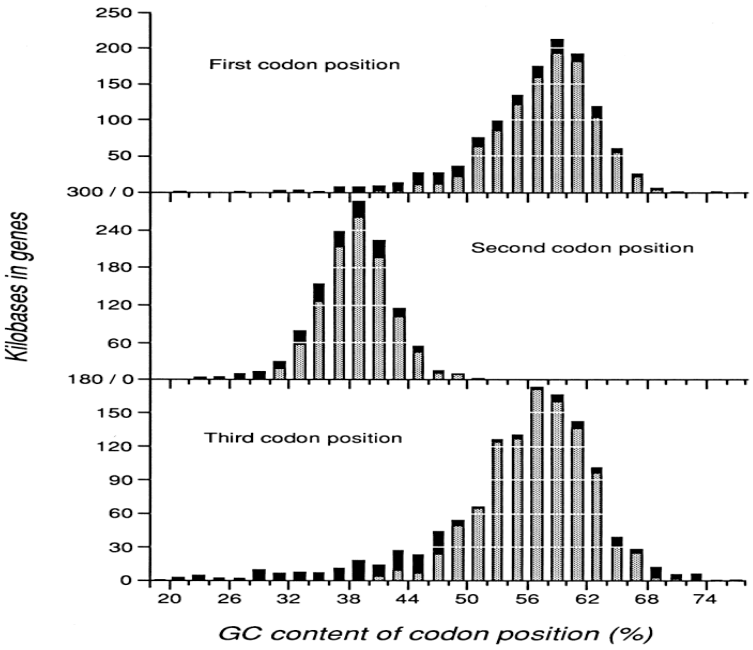

Figure 21 : Distribution du G+C pour chaque position du codon chez E. coli. Les gènes sortant d'une distribution théorique en loi normale sont considérés comme ayant été acquis récemment (barre noire). Les gènes considérés comme natifs sont indiqué avec des barres grises. D'après Lawrence et Ochman 1997.

La raison de cet appauvrissement en G+C des gènes susceptibles d'avoir été transférés n'est pas connue, mais on pourrait mettre cette observation en relation avec le fait que les génomes des bactériophages et autres éléments génétiques mobiles bactériens ont eux aussi une tendance à avoir des gènes plus pauvres en G+C que ceux de leurs hôtes (figure 22) (Blaisdel et al. 1996 ; Rocha et Danchin 2002). On constate que ce biais de composition est plus accentué chez les éléments non persistants dans les génomes, comme les phages dsADN lytiques ou les phages à ARN, tandis que les éléments tels que plasmides, transposons ou phages tempérés ont un enrichissement en A+T plus faible en comparaison de ceux de leurs hôtes. On ne connaît pas les raisons de ce biais, serait il en rapport avec une protection contre les systèmes de restriction/modification de leurs hôtes ? Ou, comme le propose Rocha et Danchin (2002) parce que l'ATP est plus abondant dans les cellules que les autres nucléotides et parce que l'ATP et l'UTP sont moins coûteux à fabriquer que le GTP ou le CTP (compétition métabolique) ?

Les gènes identifiés comme issus de transfert et les gènes des éléments génétiques mobiles partagent donc une caractéristique commune, celle d'être appauvris en G+C. Ces deux observations mises bout à bout pourrait indiquer qu'une bonne partie de tous les gènes

identifiés comme issus de transferts horizontaux sont originaires d'éléments génétiques mobiles (Daubin 2003), ce qui permettrait d'expliquer pourquoi la plupart d'entre eux n'ont pas de fonction connue et souvent pas ou peu d'homologues cellulaires (les génomes viraux sont en effet très majoritairement constitués de gènes de fonction inconnue et ayant peu ou pas d'homologues cellulaires). De plus cela indique que ces éléments génétiques seraient dans ce cas bien plus que de simples "transporteurs" de gènes, d'espèce cellulaire à espèce cellulaire, mais qu'ils seraient en fait capables d'inventer de nouvelles fonctions et de promouvoir les transferts horizontaux de ces fonctions vers leurs hôtes cellulaires.

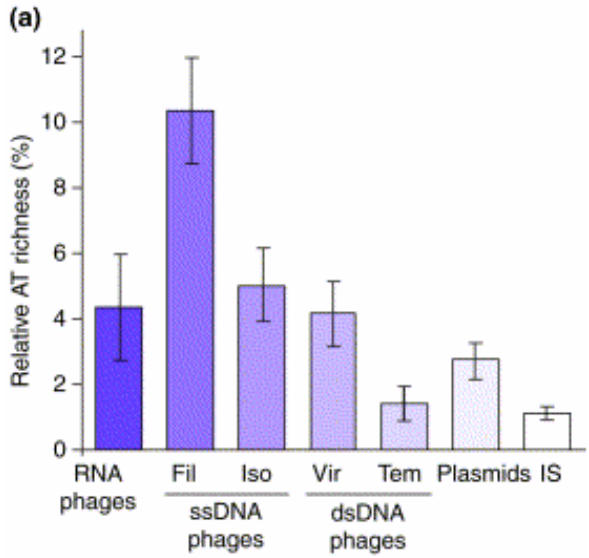

Figure 22 : Richesse relative en A+T des différentes classes d'éléments génétiques mobiles par rapport à leurs hôtes cellulaires. Fil : phages filamenteux, Iso : phage isométriques, Vir : phages virulents, Tem : phages tempérés. Extrait de Rocha et Danchin 2002.

Cette vision du rôle potentiellement important des virus au cours de l'évolution de leurs hôtes est récente et assez peu d'investigations ont été menées à ce sujet. Etant donné que les virus codent très souvent des enzymes de la réplication et du métabolisme de l'ADN, plusieurs hypothèse ont été proposées récemment postulant que beaucoup d'enzymes cellulaires informationnelles ont été inventées par les virus (Forterre 1999, Villareal et DePhilippis 2000, Takemura 2001).

CHAPITRE C

L'Evolution de l'appareil de réplication de l'ADN

I. La réplication de l'ADN cellulaire

a. Mécanismes et composants de la réplication de l'ADN au sein des trois domaines du vivant.

Tous les génomes cellulaires connus possèdent un génome ADN double brin, linéaire chez les Eucaryotes et circulaire chez les Procaryotes. Les principes généraux de la réplication de l'ADN sont très similaires au sein des trois domaines du vivant. Le processus implique (figure 23) :

- une étape d'initiation catalysée par une protéine reconnaissant spécifiquement une origine de réplication.
- un facteur de charge qui associe l'Hélicase Réplicative au complexe d'initiation précédemment formé.
- la fourche de réplication progresse grâce à l'action concertée d'une ADN Hélicase qui ouvre la double hélice, de protéines "single strand binding" qui stabilise les simples brins formés par l'Hélicase, d'une Primase qui réalise la synthèse des amorces ARN, et de 2 ADN polymérases (encore appelée "Réplicases") pour la synthèse couplée des brins "leading" et "lagging" . Des facteurs de processivité associés aux ADN polymerase permettent la synthèse de grands fragments d'ADN.
- La progression du réplisome induit au niveau de la double hélice d'ADN l'accumulation de super-tours positifs, une ADN topoisomérase est donc, en aval, nécessaire pour résoudre ce problème topologique et permettre effectivement l'ouverture progressive de la double hélice.

Si ce mécanisme général est similaire dans l'ensemble du monde cellulaire, le mode de réplication ne l'est pas. Les Eucaryotes possèdent plusieurs origines de réplication, tandis que les Bactéries et les Archéobactéries n'en possèdent qu'une seule (Myllykallio et al. 2000). De plus, beaucoup de protéines impliquées dans la réplication chez les Bactéries n'ont pas d'homologue chez les Archéobactéries et les Eucaryotes, tandis que beaucoup de composants Eucaryote sont homologues à des enzymes des Archéobactéries (figure 23).

Ces fortes différences entre l'appareil de réplication des Eucaryotes et des Archéobactéries d'une part, et des Bactéries d'autre part, sont d'autant plus surprenantes que les principaux composants de l'appareil de transcription (comme les ARN polymérases ADN dépendantes), ainsi que de nombreux composants de la machinerie de traduction, sont universellement conservés (Leipe et al. 1999). La question qui se pose naturellement est donc : comment expliquer évolutivement ces observations ?

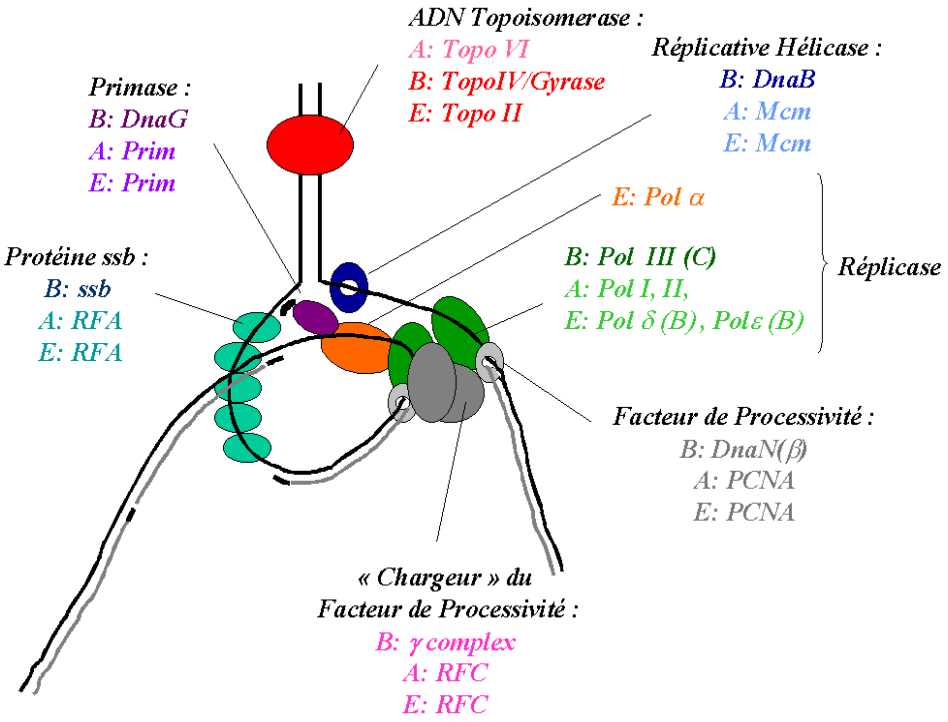

Figure 23 : Fourche de réplication universelle détaillant les différentes enzymes impliquées dans les trois domaines du vivant. Pour chaque enzyme, le nom est donné, si les enzymes sont homologues elles sont indiquées de la même couleur, A pour Archéobactérie, B pour Bactérie, et E pour Eucaryote. Pour la Réplicase, les lettres entre parenthèses indiquent la famille d'ADN polymérase correspondante. Adapté de Forterre, Filée et Myllykallio 2003, voir en annexe.

b. Les hypothèses proposées pour expliquer les profondes différences de l'appareil de réplication entre Eucaryote/Archéobactérie et Bactérie.

1) Le système de réplication de l'ADN est bel et bien homologue entre les trois domaines du vivant, mais il a divergé à tel point au niveau des séquences primaires que la reconnaissance de cette homologie est impossible par comparaison de séquences (Edgell et Doolittle 1997). En faveur de cette hypothèse on peut noter que le facteur de processivité des Eucaryotes et Archéobactéries (PCNA) et celui des Bactéries (DnaN), bien que ne possédant pas de ressemblance au niveau de la séquence primaire, possèdent une structure tridimensionnelle identique et superposable (Kelman et O'Donnell 1995). Ces composants sont donc très probablement homologues malgré leurs fortes divergences. De plus certaines enzymes a priori non-homologues possèdent en commun certains domaines protéiques. Par exemple l'initiateur de la réplication, DnaA chez les Bactéries et Cdc6/Orc1 des Archéobactéries et Eucaryotes partagent un module ATPase de la même famille (Erzberger et al. 2002). Néanmoins la plupart des autres protéines n'ont aucune similarité, à tous les niveaux. C'est particulièrement vrai pour les primases. En effet la primase bactérienne possède un domaine protéique particulier (Toprim) que l'on retrouve chez d'autres enzymes telles que des Nucléases ou les Topoisomérases I, II et VI mais que l'on ne retrouve pas chez la primase des Eucaryotes et des Archéobactéries. Ces deux enzymes, analogues fonctionnels, ne sont donc pas homologues (Leipe et al 1999). Il en est de même pour d'autres composants-clés de la réplication comme pour les ADN polymérases (Leipe et al. 1999).

2) Le LUCA avait un génome à ARN et l'ADN a été inventé deux fois séparément, une fois dans la lignée Eucaryote/Archéobactérie et une fois dans la lignée des Bactéries (Leipe et al. 1999). Cette hypothèse suppose que les Eucaryotes soient effectivement groupe frère des Archéobactéries. Il est, de plus, difficile dans ces conditions d'expliquer pourquoi certains composants informationnels comme les ARN polymérases ADN dépendantes où les facteurs de processivité (PCNA/DnaN et RFC/γ complex) sont homologues au sein des trois domaines.

3) L'appareil de réplication ancestral du génome à ADN du LUCA a été remplacé, soit chez les Bactéries, soit chez les Archéobactéries/Eucaryotes, par un nouveau système (Edgel et Doolitle 1997). Ce remplacement non-orthologue des composants de la réplication du LUCA ne résout pas le problème de l'origine évolutive de ce système

inventé secondairement. Certains imaginent que le LUCA possédait deux systèmes différents de réplication qui coexistaient, un système utilisé pour la réparation de l'ADN et un système utilisé au sens propre pour la réplication de l'ADN. Il y aurait eu perte différentielle d'un des deux systèmes dans chaque lignée (Edgell et Doolitle 1997). Le système de réparation aurait donc remplacé le système de réplication dans une des lignées. D'autres, comme Forterre (1999) ou Villareal (2000), proposent que les virus auraient été la source de ces gènes ayant remplacé dans chaque lignée des gènes cellulaires par des non-homologues viraux. En faveur de cette hypothèse, on peut noter que les génomes viraux contiennent tout ou partie de leur propre appareil de réplication qui n'est souvent que partiellement homologue à celui de leur hôte. Un tel cas de remplacement non-homologue d'une enzyme cellulaire par une enzyme virale est documenté pour l'ARN polymérase des mitochondries qui n'est pas de type bactérien (alors que les mitochondries divergent d'une α Protéobactérie) mais d'un type exclusivement viral appartenant à la classe des bactériophages T3/T7 (Figure 24)(Gray et Lang 1998 ; Moreira 2000).

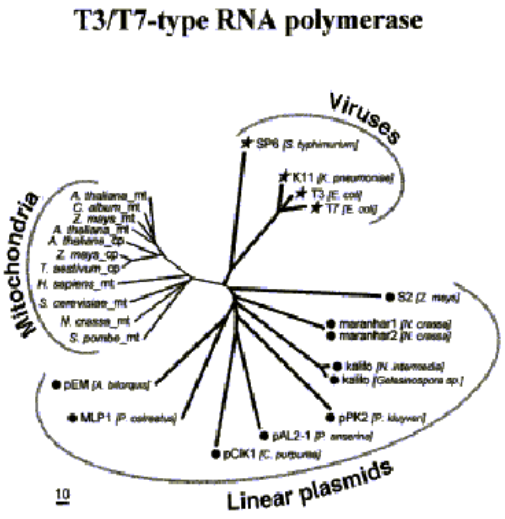

Figure 24 : Phylogénie de l'ARN polymérase de la famille T3/T7 (issu de Moreira 2000)

II. La réplication de l'ADN chez les virus

La réplication de l'ADN chez les virus procède :

- soit par un mécanisme analogue à la réplication de l'ADN dans les cellules (en utilisant la totalité de la machinerie de l'hôte comme SV40, ou en codant, au moins pour partie, une machinerie de réplication propre, comme le bactériophage T4 ou les Herpesvirus)
- soit par des mécanismes différents qui font intervenir des enzymes ne possédant pas d'homologue cellulaire, et parfois en détournant le système de réplication de l'hôte .

Les Adénovirus possèdent un système de réplication très différent de ceux de leurs hôtes Eucaryotes (De Pamphilis 1996). La réplication est continue, un brin a la fois, grâce aux déplacements du réplisome qui, lorsque le premier brin est synthétisé, "saute" sur l'autre brin pour le répliquer. Il n'utilise pas une Primase pour la synthèse des amorces ARN. C'est une ADN polymérase particulière dite "protein-primed", codée par le virus, qui catalyse la formation d'une liaison phosphodiester entre le premier nucléotide de la nouvelle chaîne d'ADN (une cytosine) et le résidu sérine d'une protéine accessoire codée elle aussi par le virus et sans homologue cellulaire. Enfin l'élongation n'implique que 2 protéines différentes (l'ADN polymérase qui possède aussi une activité hélicase et une protéine ssb non homologue aux protéines ssb cellulaires). Il s'agit d'un mécanisme très simple, efficace et complètement distinct de celui de l'hôte.

Les Parvovirus utilisent eux aussi un mécanisme d'initiation de la réplication original. Leurs génomes sont constitués d'ADN simple chaîne et possèdent des terminus complémentaires en épingle à cheveux ("brin plus"). C'est donc l'extrémité simple brin 3' qui joue le rôle d'amorce de la réplication, sans avoir besoin d'une Primase. L'élongation du brin complémentaire ("brin plus") implique ensuite la plupart des enzymes cellulaires responsables de la synthèse de l'ADN nucléaire. Cette forme double brin est ensuite répliquée d'une manière continue, un brin a la fois, suivant un mécanisme original "en cercle roulant" (Figure 25). Lorsque le brin plus est synthétisé, une nucléase codée par le virus (protéine "Rep") le clive d'une manière site-spécifique , Rep se fixe aux terminus 5' d'une manière covalente et l'extrémité 3'OH libre est utilisée comme amorce pour la synthèse d'un nouveau brin plus naissant qui remplace le brin plus parental. Le brin plus néo-synthétisé est clivé par Rep ce qui permet son relarguage

et le commencement d'un nouveau cycle. La réplication "en cercle roulant" est partagée par de nombreux phages (Kornberg et Baker 1992), des virus (De Pamphilis 1996) et des plasmides ssADN de Bactérie (Ilyina et Koonin 1992) et d'Archéobactérie (Erauso et al. 1996).

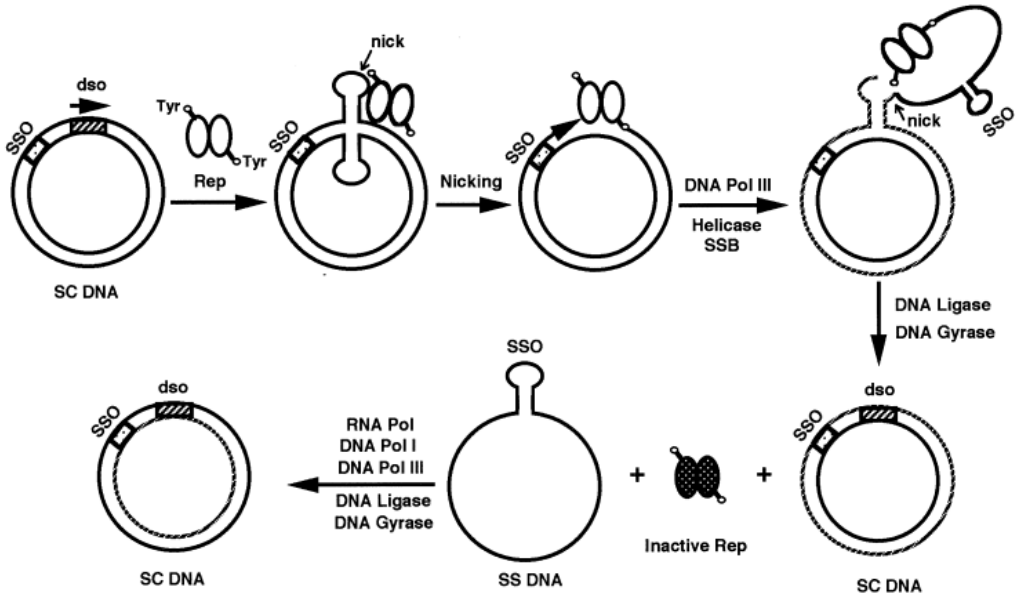

Figure 25 : Modèle général pour la réplication en cercle roulant. SSO : origine de réplication simple brin, dso : origine de réplication double brin. Voir texte pour les détails. Tiré de Khan 2000.

Les virus qui utilisent un mécanisme de réplication analogue à celui des cellules codent aussi pour leurs propres enzymes, qui ne sont pas toujours homologues aux composants cellulaires correspondants. C'est par exemple le cas des Hélicases de la super famille III chez beaucoup de virus dsADN eucaryotes (Iyer et al. 2001), les 3 enzymes du complexe Primase/Hélicase chez les Herpesvirus (Dracheva et al. 1995), la "single strand binding" protéine de ICP8 des Herpesvirus (De Pamphilis 1996) ou encore l'ADN polymérase/Primase de la famille E du plasmide pRN2 de *Sulfolobus islandicus* (Lipps et al. 2003). Il est aussi très probable qu'un grand nombre d'enzymes virales impliquées dans la réplication des génomes viraux et sans homologue cellulaire restent à découvrir. Par exemple le génome de 280kb du bactériophage phiKF de *Pseudomonas aeruginosa* ne contient aucune protéine connue pour être impliquée

dans la réplication de l'ADN ce qui suggère l'existence de protéines de la réplication très divergentes de celles connues jusqu'à présent (Mesyanzhinov et al. 2002).

L'existence d'un tel réservoir d'enzymes de la réplication chez les virus, parfois non-homologues de celles des cellules, suggère que les virus pourraient avoir été à la source de certaines fonctions, recrutées secondairement par les cellules et remplaçant ou non un composant cellulaire. Forterre (1999) a proposé que beaucoup d'enzymes de la réplication chez les Bactéries avaient été déplacées par des composants non-homologues originaires de bactériophages. Toutefois Moreira (2000) a montré que le transfert horizontal de ces gènes pouvaient être polarisé dans l'autre sens : les bactériophages "recrutant" des fonctions cellulaires. Chez les Eucaryotes, plusieurs ADN polymérases pourraient avoir une origine virale : l'ADN polymérase δ (famille B) auraient été héritée d'un virus dsADN apparenté aux Phycodnavirus/HerpesVirus (Villareal et De Filippis 2000) et l'ADN polymérase α proviendrait d'un Poxvirus (Takemura 2001). Ce dernier auteur va même plus loin en proposant que le noyau Eucaryote proviendrait de la symbiose d'un Poxvirus et d'une Archéobactérie ce qui permettrait de rendre compte de la forte ressemblance au niveau des composants informationnels entre Archéobactéries et Eucaryote. Cette dernière hypothèse a été simultanément proposée par Bell (2001). Enfin une hypothèse a été proposée récemment par Forterre (2002) qui postule que les virus auraient inventé l'ADN lui même a partir d'un monde cellulaire ARN, afin de se protéger des enzymes de dégradation spécifique de l'ARN exogène (RNAse) mises en place par les cellules. Cette hypothèse implique que les virus auraient inventé à la fois, tout ou partie d'un appareil de réplication, ainsi qu'une voie métabolique permettant la synthèse des composants de l'ADN.

III. Origine et évolution du métabolisme terminal de l'ADN

L'ADN peut être considéré comme une forme modifiée de l'ARN. Les formes de vie actuelles produisent en effet les bases de l'ADN par réduction des ribonucléotides di- ou tri-phosphate. Associés au fait que les ARN sont capables d'activité auto-catalytique (ribozyme), ces arguments favorisent l'idée selon laquelle l'ADN est apparu postérieurement à l'ARN.

Les principales étapes de la formation des déoxynucléotides a partir des déoxyribonucléotides sont schématisées figure 26. Les premières molécules d'ADN contenaient probablement de l'uracile et non de la thymine. En effet, le TTP n'existant pas dans les cellules, les enzymes qui réduisent les ribonucléotides (Ribonucléotide Réductase) produisent du dUTP/dUDP à partir de l'UTP/UDP. Le dTTP est produit à partir du dUTP et non pas du TTP. L'ADN-T serait donc apparu après l'ADN-U. On peut noter que l'ADN-U existe dans le génome de certains virus actuels (Takahashi et Marmur 1963) : ce pourrait être une rémanence d'une situation ancestrale. Le dTTP est produit en trois étapes :

- Formation de dUMP, soit par déphosphorylation du dUTP par une dUTPase, soit par déamination du dCMP par la dCMP déaminase.
- Méthylation du dUMP en dTMP par la Thymidylate Synthase.
- Phosphorylation du dTMP en dTTP par la Thymidine Kinase.

Cette transition ARN vers ADN-U puis ADN-T pose toutefois le problème de l'origine des enzymes nécessaires. Forterre (2002) propose que ces enzymes ont été inventées par les virus pour se protéger des systèmes de défense de l'hôte basés sur des Rnases. Néanmoins, les plus gros génomes viraux ARN connus à ce jour font à peine 30kb [Coronavirus, (Atkins 1993)], et généralement beaucoup moins. Ces petits génomes ne sont notoirement pas assez grands pour coder le stock d'enzymes minimal nécessaire la transition ARN vers ADN (au minimum une Ribonucléotide réductase, une ADN polymérase et une ADN hélicase)(Poole et al. 2000). Deux raisons expliquent les faibles capacités de codages des génomes ARN :

- Les enzymes qui répliquent ces génomes ne possèdent pas de capacité de correction des erreurs, ce qui entraîne des taux de mutation très élevés. Plus le génome est grand, plus le nombre de mutations augmente.

- La présence d'un groupement hydroxyle en 2' sur le ribose (figure 27) susceptible d'attaquer la liaison phosphodiester de la chaîne et d'entraîner une cassure de brin.

On pourrait aussi imaginer que si les virus ARN actuels sont si petits c'est parce qu'ils ont évolué par régression à partir d'élément ayant possédé des génomes à ARN beaucoup plus gros. Cette transition aurait été sélectivement avantagée par la baisse des taux de mutation et de cassure des brins chez les virus ARN possédant des petits génomes. Inversement, les cellules primitives ayant possédé un génome ARN aurait "choisi" une voie évolutive différente, celle de posséder des génomes de plus grande taille, codant pour un plus grand nombre de fonctions.

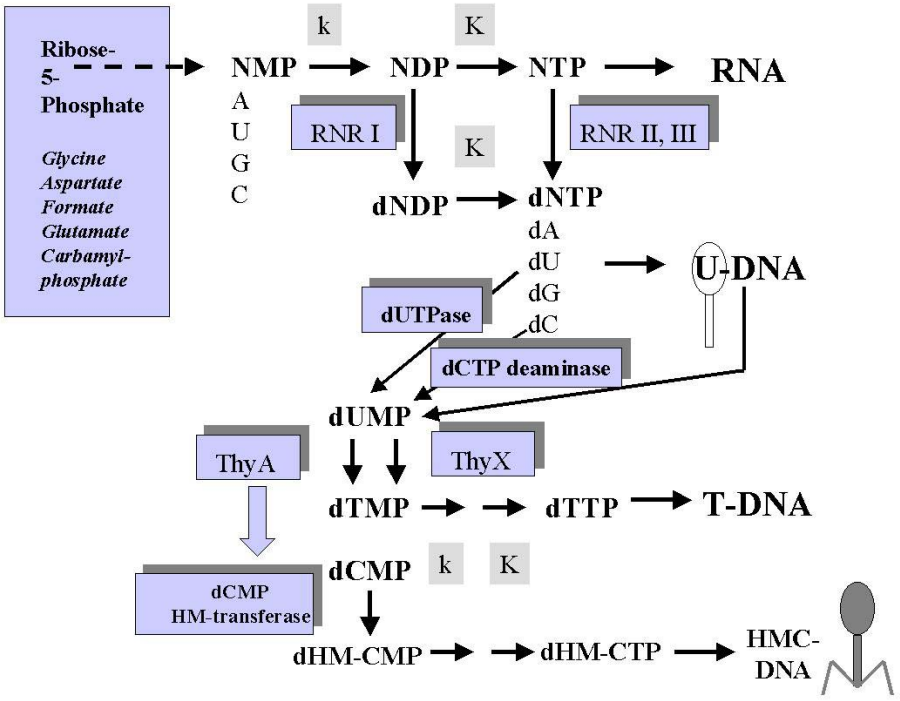

Figure 26 : Voie de biosynthèse des composants de l'ADN et de l'ARN. Le NMP est converti en NTP par la NMP Kinase (k) et la NDP kinase (K). Ces composants sont réduit par les Ribonucléotides Réductases (RNR I, II et III) pour former du dNTP. Le DTTP est produit en trois étapes impliquant dUTPase, dCTP déaminase et Thymidylate Synthase (ThyX et ThyA). Certains bactériophages comme T4 présentent des génomes modifiés au niveau des cytosines par hydroxyméthylation catalysée par une dCMP Hydroxyméthyl transférase (homologue à ThyA). Adapté de Forterre, Filée, Myllykallio (2003), voie en annexe.

Pour permettre à un génome ARN (viral ou cellulaire) de coder la machinerie nécessaire pour accomplir la transition vers un monde ADN-U, il faut imaginer un moyen de stabiliser l'ARN. Il a été proposé que la méthylation des ARN sur l'oxygène en 2' du ribose (figure 26) aurait accompli cette tâche (Poole et al. 2000). La méthylation des ARN en 2' sur le ribose est un processus ubiquiste au sein du vivant, en particulier au niveau des ARN ribosomiques. Il est à noter que certains virus dsADN Eucaryote comme les Poxvirus (Hodel et al. 1996) ou les Baculovirus (Wu et Guarino 2003), et certains virus ARN comme les Flavivirus ou les Alphavirus (Feder et al. 2003) codent eux aussi pour leur propre méthyl-transférase (les méthyl-transférases des virus dsADN étant homologue de méthyl-transférases cellulaires). L'hypothèse proposée par Poole et collaborateurs (2000) pour expliquer l'avantage sélectif du passage d'un monde méthyl-ARN à un monde U-ADN est la suivante : la méthylation du ribose des ARN est un mécanisme post-réplicatif, alors que les Ribonucléotides réductases réduisent le 2' OH du ribose en 2' H d'une manière pré-réplicative, supprimant l'étape où l'ARN qui n'a pas encore été modifié peut potentiellement s'auto-cliver.

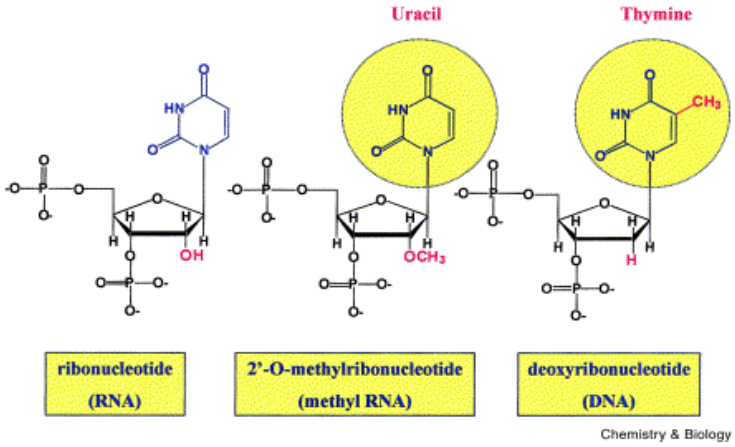

Figure 27 : Formule chimique des nucléotides tels qu'ils apparaissent de nos jours dans les chaînes d'acide nucléique. La différence entre l'ARN, le 2'-O-méthyl ARN et l'ADN consiste en la nature du groupe en position 2' du ribose (indiqué en rouge).

Toutefois, les explications basées sur la stabilité pour rendre compte du passage ARN ou méthyl-ARN vers ADN ont un inconvénient : elles ne procurent pas d'avantage sélectif à court terme au premier génome ADN formé qui est en compétition avec une population d'éléments à génome ARN ou méthyl-ARN. Forterre (2002) propose que les virus auraient

inventé l'ADN pour se protéger des Rnases de l'hôte ce qui leur aurait procuré un avantage sélectif immédiat. Le système aurait été secondairement recruté par les cellules, par exemple suivant le même schéma que pour les systèmes R/M de l'ADN (voir chapitre B partie V.). Au crédit de cette hypothèse on peut noter que certains bactériophages comme les T-pairs présentent des génomes dont les bases sont effectivement modifiées au niveau des cytosines par ajout d'un radical hydroxyméthyl. Ceci afin de se protéger des systèmes de restriction de leurs hôtes. L'enzyme qui catalyse cette réaction (la dCMP hydroxyméthyl transférase) modifie les cytosines avant leur incorporation dans l'ADN. De plus cette enzyme est homologue à la Thymidylate Synthase ThyA qui modifie le dUMP en dTMP (Song et al. 1999). De plus, les hypothèses de Poole et collaborateurs (2000) et Forterre (2002) sont parfaitement complémentaires, on pourrait imaginer que les cellules auraient inventé le méthyl-ARN leur permettant de disposer d'un génome de plus grande taille, pour coder un nombre de fonctions plus élevé. Les virus au cours de leur "course aux armements" pour échapper aux systèmes de défense de leurs hôtes auraient pour certains d'entre eux développé différents systèmes pour modifier leurs génomes ARN vers des génomes U-ADN ou HMC-ADN. Les cellules ayant acquis secondairement les gènes viraux nécessaires à la synthèse d'ADN par transferts horizontaux. Enfin on pourrait aussi imaginer le scénario inverse de celui proposé par Forterre : les cellules modifiant leur génome ARN pour échapper aux enzymes de dégradation des virus. Les virus auraient, pour certains d'entre eux, recruté secondairement ces fonctions cellulaires. Ce qui permettrait d'expliquer pourquoi il existe encore aujourd'hui beaucoup de virus à génome ARN et aucun génome cellulaire à ARN.

Finalement, une dernière question se pose : pourquoi est-on passé d'un ADN-U vers un ADN-T chez les cellules et certains virus ? La déamination spontanée des cytosines en uracile provoquant beaucoup de mutations délétères, un génome ADN-T capable de détecter ces uraciles et de les éliminer via un système de réparation de l'ADN adéquat bénéficie alors d'un avantage sélectif majeur (Poole et al. 2001). Cette hypothèse expliquerait donc l'évolution d'un ADN-U vers un ADN-T. Alternativement, on pourrait aussi penser que l'ADN-T aie été inventé primitivement par les virus pour se protéger des Dnases-U cellulaires, puis acquis par les cellules pour résoudre le problème de la déamination spontanée des cytosines.

IV. Conclusions

L'appareil de réplication de l'ADN est universel, il est présent chez tous les organismes cellulaires et de nombreux virus codent tout ou partie de leur propre système. Au cours de ce chapitre, nous avons vu qu'il n'était pas conservé au sein des trois domaines du vivant. Différentes hypothèses ont été énoncées précédemment pour expliquer cette observation. L'une d'elle, proposée simultanément par Patrick Forterre (1999) et Luis Villarreal (2000), postule que les virus ont inventé de nombreux composants de la réplication, et que ces gènes d'origine virale ont remplacé certains composants d'origine cellulaire. Forterre (2002) allant même plus loin en proposant que les virus eux même ont primitivement "inventé" l'ADN et sa réplication, les cellules auraient recruté ces fonctions dans un deuxième temps pour accomplir la transition d'un monde cellulaire ARN vers un monde ADN. Ces deux hypothèses sont principalement soutenues par le fait que les génomes viraux codent pour un grand nombre d'enzymes impliquées dans le métabolisme et la réplication de l'ADN. De plus, une partie de ces enzymes n'ont aucun homologue cellulaire et peuvent donc être considérées comme ayant effectivement été inventées par les virus.

Toutefois ces observations ne sont pas validées d'un point de vue phylogénétique. En effet les virus pourraient très bien avoir acquis ces gènes en provenance des cellules et non l'inverse. Il est donc nécessaire de reconstruire l'histoire évolutive de tous ces gènes pour trancher entre ces deux hypothèses. L'étude des composants de la réplication chez les virus pourraient aussi nous apporter des éléments de réflexion sur l'évolution des virus eux-mêmes.

RESULTATS

Dans cette partie, le lecteur trouvera l'analyse phylogénétique d'un grand nombre de gènes dont les produits sont impliqués dans la réplication et le métabolisme de l'ADN. Plutôt que de faire un catalogue exhaustif de ces analyses, nous avons pris le parti de diviser la présentation des résultats en différents chapitres qui illustrent chacun un aspect de l'évolution des enzymes informationnelles. Certains de ces résultats ont fait l'objet de publications qui précisent et complètent les informations données dans le chapitre. Ces articles sont présentés à la fin de chaques chapitres.

Nous parlerons d'abord des ADN polymérases et des difficultés méthodologiques à polariser les transferts de gènes entre virus et cellules.

Nous évoquerons ensuite le cas des ADN toposiomérases et de leurs histoires évolutives complexes où l'on ne retrouve pas l'habituelle et profonde différence entre l'appareil de réplication des Archéobactéries/Eucaryotes d'une part et celui des Bactéries d'autres part.

Parmi toutes les autres enzymes de la réplication, nous avons aussi identifié d'autres cas où des gènes viraux auraient potentiellement été recrutés par les cellules pour accomplir des fonctions cellulaires. Un chapitre sera donc consacré à l'histoire évolutive de ces enzymes.

Enfin, dans une dernière partie nous nous intéresserons à l'évolution des enzymes nécessaire à la transition d'un monde à ARN vers un monde à ADN et nous montrerons que les virus y ont joué un rôle important, au moins à petite échelle évolutive.

CHAPITRE D

Les ADN polymérases

a. Généralités.

Les ADN polymérases forment un vaste groupe d'enzymes constitué d'au moins 7 familles qui ne sont pas homologues entre elles. Aucune de ces 7 familles n'est universellement conservée dans le monde vivant actuel, et de très nombreux génomes viraux ou plasmidiques codent pour leur propre ADN polymérase. Ainsi, la question des relations évolutives entre les ADN polymérases cellulaires et virales se pose naturellement.
Certaines familles ont une distribution phylogénétique réduite : la famille D est uniquement présente chez les EuryArchéobactéries, la famille C est présente chez la totalité des génomes bactériens séquencés et chez quelques virus et plasmides, alors que la famille E est seulement représentée par une ADN polymérase codée par le plasmides pRN1 de l'Archéobactérie *Sulfolobus islandicus* (Lipps et al. 2003). D'autres familles ont une répartition phylogénétique beaucoup plus vaste, comme les familles A et Y, qui sont présentes chez de nombreuses Bactéries, Eucaryotes et virus, ou encore la famille B qui est présente chez les Eucaryotes, les Archéobactéries, de nombreux virus et quelques gamma Protéobactéries.

Du point de vue fonctionnel, les ADN polymérases illustrent parfaitement la différence générale de l'appareil de réplication des Bactéries en comparaison de celui des Archéobactéries et des Eucaryotes. En effet les Bactéries utilisent une ADN polymérase de la famille C pour répliquer leur génome, alors que les Eucaryotes et les Archéobactéries utilisent des ADN polymérases de la famille B (et de la famille D pour les EuryArchéobactéries). Les virus codent et utilisent une large gamme d'ADN polymérases. L'objectif principal de la publication ci-après étaient de clarifier les relations phylogénétiques entre toutes ces ADN polymérases et, en particulier, d'évaluer la contribution des virus concernant le remplacement de gènes cellulaires par des gènes viraux.

b. La polarisation des transferts

L'analyse phylogénétique de toutes les familles démontre que l'échange de gènes entre virus et cellule est un phénomène fréquent et qu'il se produit dans les deux sens. Toutefois la détermination de la polarité du transfert d'un gène pose problème. Est-ce les cellules qui acquièrent un gène d'origine virale ou plutôt l'inverse ? En effet, hormis le cas des ADN polymérases de la famille Y, les enzymes virales ne sont pas positionnées dans les arbres à proximité de celles de leurs hôtes. On peut donc exclure un transfert "récent" de gènes

polarisé de la cellule vers le virus, mais on ne peut pas rejeter l'hypothèse d'un transfert "ancien" des cellules vers les virus, suivi ou non d'une accélération de la vitesse d'évolution des séquences virales, les rejetant d'une manière très distante de leurs hôtes dans les phylogénies. Nous avons donc besoin d'un critère a priori pour trancher entre ces deux hypothèses. Nous pouvons d'abord utiliser le critère de la paralogie si la famille a connu des duplications de gènes (Figure 28).

Si un gène est présent en au moins deux exemplaires dans un génome donné, soit il s'agit d'une duplication de gène, soit l'une des copies est issue d'un transfert horizontal. Dans le cas d'une duplication, on s'attend à ce que les deux gènes soient positionnés en groupes frères. Ce cas de figure n'est pas toujours retrouvé. L'hypothèse la plus parcimonieuse implique dans ce cas un transfert horizontal, même si l'on ne peut exclure de multiples pertes sélectives d'un des deux gènes. Si une séquence virale se positionne à la base d'un de ces deux types de gènes, il est donc très probable que l'organisme "donneur" soit le virus, ainsi peut on polariser le sens du transfert.

Ce cas de figure est illustré avec la famille B des ADN polymérases. Simultanément à notre travail, Villarreal et DePhillipis (2000) publiaient la phylogénie de cette famille en affirmant que le "paralogue" delta était originaire d'un virus. Or, la phylogénie complète de cette famille est très mal résolue, et l'on ne peut pas dire avec confiance que les paralogues sont phylogénétiquement très éloignés. Toutefois, des analyses avec des échantillons plus réduits positionnent en effet plusieurs groupes de virus (Phycodnavirus, Herpesvirus et Iridovirus) à la base du groupe "delta" (voir article V). Il est donc en effet possible que l'ADN polymérase delta des Eucaryotes soit d'origine virale, mais la polarisation inverse ne peut être exclue puisque l'on pourrait imaginer une récupération ancienne du gène par l'ancêtre de tous ces virus. Par contre, au sein de cette famille B, un meilleur exemple de transfert de virus vers cellules, à plus petite échelle évolutive, nous est donné avec une des deux ADN polymérases d'*Halobacterium salinarum* ("paralogue" B1). En effet, non seulement ce paralogue est très éloigné phylogénétiquement du paralogue B2, mais il est également très distant des autres séquences d'Archéobactéries. Enfin, un groupe de virus d'Archéobactérie Halophile représenté par les éléments HF1/HF2, se positionne en groupe frère du gène B1. On se trouve donc typiquement dans la situation énoncée figure 28 : l'ADN polymérase B1 d'*Halobacterium salinarum* serait donc très probablement d'origine virale.

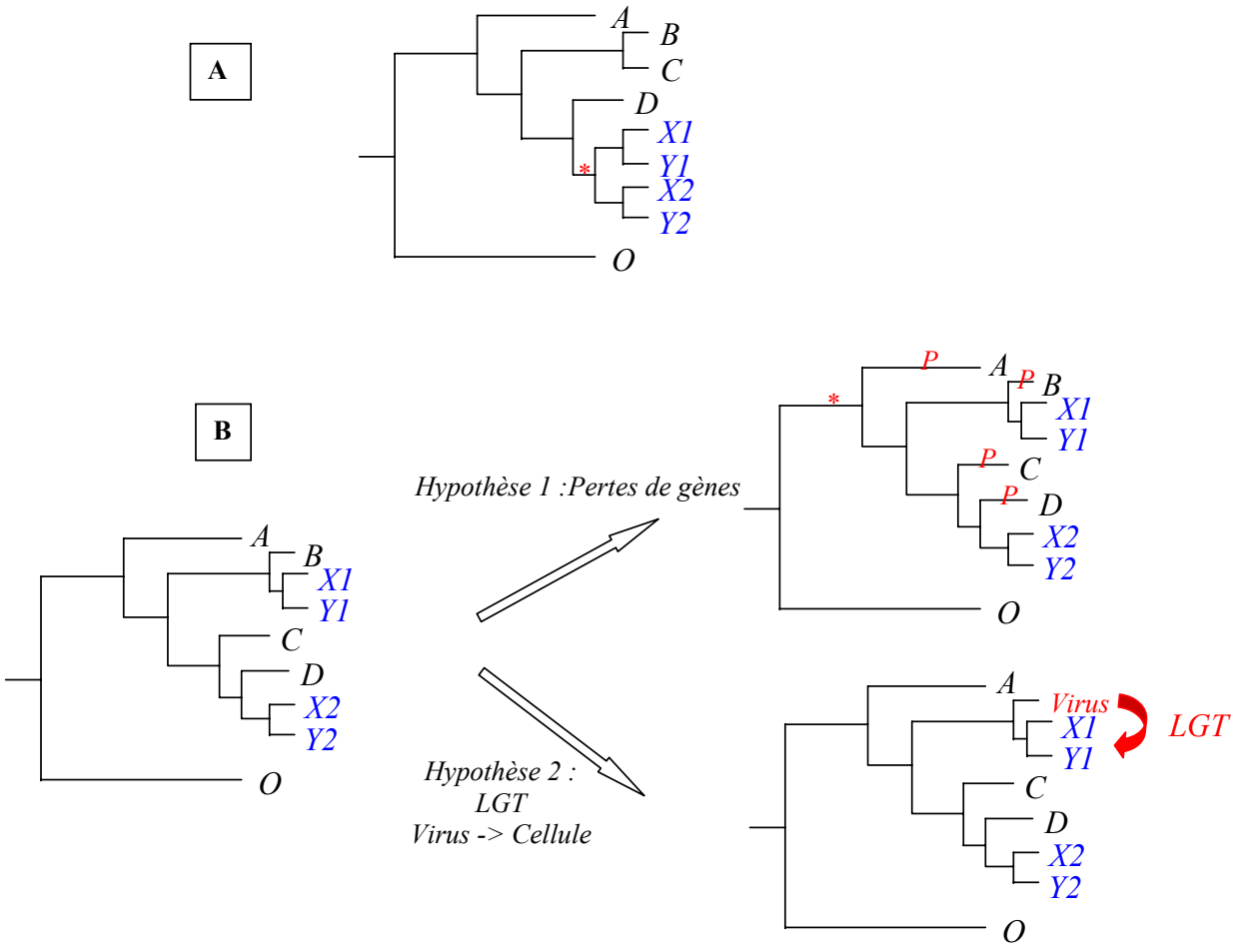

Figure 28: Polarisation du transfert de gène virus vers cellule en utilisant le critère de la paralogie.

A. La phylogénie montre une duplication symbolisée par un astérisque rouge chez l'ancêtre de X et Y ; A, B, C et D représentant les autres taxons de la phylogénies et O le groupe extérieur permettant l'enracinement. Les paralogues 1 et 2 se placent en groupes frères l'un de l'autre.

B. Lorseque deux "paralogues" coexistent dans un même génome, on peut parfois observer que les gènes de type 1 et les gènes de type 2 paraissent phylogénétiquement éloignés. Ce cas de figure peut s'expliquer de 2 façons. Dans l'hypothèse 1, une duplication ancestrale chez tous les taxa (astérisque rouge) est suivie de nombreuses pertes sélectives de l'un ou l'autre gène (indiqué par la lettre P sur le shéma). Ou alors, hypothèse 2, un transfert horizontal de l'un des deux gènes s'est produit chez l'ancêtre de X et Y; si une séquence virale se place à la base de l'un de ces 2 gènes (remplaçons le taxon B par un virus), on possède un bon critère pour polariser le transfert du virus vers la cellule et non l'inverse. LGT : transfert horizontal de gène.

Nous pouvons aussi dans certains cas utiliser nos connaissances *a priori* de l'évolution pour polariser les transferts entre cellules et virus. Par exemple, le fait que la mitochondrie résulte de l'endosymbiose d'une alpha-Protéobactérie ; dans les phylogénies, on s'attend donc à trouver les séquences mitochondriales à proximité des séquences des Protéobactéries, et tout du moins, au sein d'un groupe de Bactéries.

On sait que les mitochondries utilisent une ADN polymérase de la famille A pour répliquer leur génome, ce qui est en soi déjà surprenant puisque les Bactéries utilisent une ADN polymérase de la famille C. La phylogénie de la famille A des ADN polymérases (voir article I) démontre que les séquences des mitochondries sont phylogénétiquement éloignées des séquences des Bactéries et qu'elles sont plutôt apparentées à un groupe de virus appartenant au groupe des phages de type T3/T7. Il est donc possible que l'ADN polymérase des mitochondries soit d'origine virale, en remplacement non-homologue de la réplicase Bactérienne ancestrale de la famille C. On se trouverait donc dans une situation identique à celle de l'ARN polymérase des mitochondries qui est aussi originaire d'un phage de type T3/T7.

c. Conclusion

Nos analyses phylogénétiques des différentes familles d'ADN polymérase montrent que les virus semblent avoir joué un rôle qui n'est pas négligeable au cours de l'évolution de ces enzymes. En particulier, en "inventant" des gènes codant pour cette fonction (famille E), et en permettant aux cellules d'acquérir ces gènes d'origine virale (famille A et B). Toutefois, l'interprétation des arbresatteint des limites quant à la polarisation des transferts de gènes entre virus et cellule : les transferts récents sont assez facilement polarisables, tandis que des évènements anciens ne sont polarisables que si l'on dispose d'un critère a priori.

Article I

Evolution of DNA polymerase families : evidences for multiple gene exchange between cellular and viral proteins.

Filee J, Forterre P,Sen-Lin T, Laurent J.

J. Mol. Evol. (2002), 54:763-773

CHAPITRE E

Les ADN topoisomérases

La molécule d'ADN est présente chez tous les organismes cellulaires sous la forme d'une double hélice subissant différents types de contraintes topologiques. Beaucoup de processus biologiques, comme la réplication ou la transcription, nécessitent le "déroulement" de la double hélice. L'évolution a retenu différentes enzymes capables de résoudre ces contraintes : les ADN Topoisomérases. Au sein du monde vivant, il existe deux types d'ADN Topoisomérases qui se différencient par la nature de la cassure de brin transitoire que ces enzymes catalysent. Les ADN Topoisomérase de type I provoquent une cassure simple brin tandis que les enzymes du type II introduisent une cassure double brin.

a. Les ADN topoisomérases de type I

Il existe deux familles d'ADN Topoisomérase I qui ne sont probablement pas liées évolutivement si l'on s'en tient aux ressemblance de séquences primaires. Mécanistiquement, les enzymes de la famille IA se lient à l'ADN d'une manière covalente en 5'-OH tandis que les enzymes de la famille IB se lient au 3'–phosphate. La famille IA possède une répartition phylogénétique très vaste, puisqu'elle est présente chez la plupart des Archéobactéries, des Eucaryotes et des Bactéries dont les génomes ont été séquencés ainsi que chez de nombreux plasmides et bactériophages. De plus, la Reverse Gyrase, qui est présente chez tout les hyperthermophiles (voir chapitre A, partie III), est composée en C-terminal d'un module "topoisomérase" homologue à la Topoisomérase IA. La phylogénie de cette famille n'indique pas de transferts de gènes entre les différents domaines cellulaires, mais indique de multiples transferts de gènes entre les bactériophages/plasmides et les Bactéries. Il s'agit à chaque fois d'acquisitions de gènes cellulaires par les virus (Moreira 2000).

La famille IB est une famille de molécules initialement découvertes chez les Eucaryotes. Une recherche de type BLAST (Altschul et al. 1990) avec la séquence *d'Homo Sapiens* permet de rapatrier des séquences homologues chez un grand nombre d'Eucaryotes. On retrouve aussi le gène avec des valeurs *E* aux limites du seuil significatif (10^{-3}) chez les Poxvirus puis chez quelques Bactéries appartenant à des phylums très divergents (Krogh et Shuman 2002). On remarquera que l'enzyme Eucaryote est beaucoup plus grande que l'enzyme des Poxvirus et des Bactéries (800 acides aminé en moyenne contre environ 350 chez les Poxvirus et les Bactéries). On peut enfin noter qu'une Topoisomérase IB de 980 acides aminés a aussi été isolée de l'Archéobactérie *Methanopyrus kandleri* (Slesarev et al. 1993), la séquence primaire de la protéine n'étant que très peu ressemblante avec les autres Topoisomérase IB.

La famille IB est malheureusement trop divergente au niveau de la séquence primaire pour que l'on puisse effectuer la phylogénie, mais sa répartition phylogénétique indique très probablement de nombreux évènements de pertes de gènes et/ou de transferts horizontaux de gènes chez les Bactéries, par exemple via des remplacement non homologues de la Topoisomérase IB par la Topoisomérase IA. De plus la faible ressemblance et la grande différence de taille entre les gènes Eucaryote et des Poxvirus excluent des échanges de gènes récents entre ces différents organismes.
Enfin on notera que la répartition phylogénétique des deux familles d'ADN Topoisomérase I ne suivent pas la traditionnelle dichotomie de l'appareil de réplication entre Archébacterie/Eucaryote d'une part et Bactéries d'autre part. En effet la famille IA est largement présente au sein des trois domaines du vivant et la famille IB est présente chez les Bactéries et les Eucaryotes.

b. Les ADN Topoisomérase de type II.

Les Topoisomérases de type II appartiennent à deux familles non-homologues. La famille IIA est présente chez les Eucaryotes, chez les Bactéries et chez quelques Archéobactéries. On trouve aussi cette enzyme chez quelques virus Eucaryotes ainsi que chez les bactériophages apparentés à T4. Chez les Bactéries, l'enzyme est constituée de deux sous-unités et chez les bactériophages de deux ou trois sous-unité suivant l'élément. Chez les Eucaryotes et leurs virus les deux gènes sont par contre fusionnés. La famille IIB est constituée de 2 sous-unités assemblées en hétérotétramère de type A_2B_2. Les sous-unités A et B sont présentes chez un grand nombre d'Archéobactéries, tandis que chez les Eucaryotes seule la sous unité A est largement représentée, la sous-unité B n'étant présente que chez les Plantes.

Les phylogénies de ces deux familles sont présentées dans l'article présenté à la fin de ce chapitre.
La phylogénie de la sous unité B de la Topoisomérase IIB est congruente avec la phylogénie basée sur l'ADN ribosomal. Il en est de même pour la sous-unité A avec la dichotomie Crenote/Euryote, la présence de cette sous-unité chez les Plantes s'expliquant soit par des pertes de gènes chez tous les autres Eucaryotes, soit par un transfert horizontal ancien chez l'ancêtre des Plantes en provenance d'un génome d'Archéobactérie.
La phylogénie de la Topoisomérase IIA est plus complexe. On observe une duplication ancienne du gène chez les Bactéries. La polyphilie marquée des Archéobactéries dans cet

arbre indique très probablement des événements de transferts de gènes, multiples et indépendant dans au moins trois lignées/espèces d'Archéobactérie (*Halobacterium sp.*, *Methanosarcina barkeri* et les Thermoplasmatales). On notera que, dans le phylum des Thermoplasmatales, les deux sous-unités de la Topoisomérase IIB sont absentes, cette observation allant dans le sens du remplacement non-homologue de la Topoisomérase IIB par une Topoisomérase IIA d'origine bactérienne. Enfin, certains virus Eucaryotes occupent la base de l'arbre du sous-ensemble des Eucaryotes, cette position rend toute interprétation évolutive délicate quant à la polarisation d'éventuels transferts anciens entre virus et l'ancêtre des Eucaryotes. Néanmoins cette position basale ne permet pas de rejeter l'hypothèse selon laquelle le gène Eucaryote serait d'origine virale. Ces résultats indiquent que la répartition phylogénétique de la famille II des ADN Topoisomérases est atypique en comparaison de la répartition phylogénétique de la plupart des enzymes de la réplication : la Topoisomérase IIA était très probablement présente chez le LCA des Eucaryotes et des Bactéries tandis que les deux sous-unités de la Topoisomérase IIB étaient très probablement présentes chez le LCA des Archéobactéries. La sous-unité A de la Topoisomérase IIB était sans doute aussi présente chez le LCA des Eucaryotes, la sous-unité B ayant soit été acquise par les Plantes par transfert en provenance des Archéobactéries soit était présente chez l'ancêtre et perdues dans la plupart des phyla, a l'exception des Plantes.

c. Conclusion

Les ADN topoisomérases, qu'elles soient de type I ou de type II, ont été "inventées" chacune au moins deux fois indépendamment. La répartition phylogénétique de ces 4 familles ne suit pas l'habituelle répartition des gènes de la réplication, où, généralement les protéines d'Archéobactéries et d'Eucaryotes sont homologues, à la différence de celle des Bactéries. Aucun scénario évolutif simple ne permet d'expliquer ces observations car il semblerait que de nombreux évènements de duplications de gènes, de pertes de gènes ou de transferts horizontaux suivis ou non de remplacement homologue ou non homologue ont émaillé l'histoire évolutive de ces familles. Le fait que de nombreux génomes viraux codent une ou plusieurs ADN Topoisomérases suggère que ces éléments ont pu jouer un rôle important au cours de l'histoire évolutive de ces enzymes. En particulier, l'éloignement phylogénétique des virus et des bactériophages par rapport à leurs hôtes dans la phylogénie de la famille IIA pourrait indiquer que le gène a une origine virale chez les Eucaryotes et pourquoi pas chez les

Bactéries (le LCA des Bactéries n'avaient donc, dans ce cas, ni de Topoisomérase IIA ni de Topoisomérase IIB !)

Article II

Phylogenomics of type II DNA topoisomerases

Gadelle D,Filee J*, Buhler C, Forterre P*

BioEssay (2003) 25 :232-242

CHAPITRE F

Autres protéines impliquées dans la réplication de l'ADN

Beaucoup d'autres enzymes impliquées dans la réplication de l'ADN sont codées par des génomes viraux. Parmi toutes ces enzymes nous avons identifié deux cas où des gènes viraux ont potentiellement été acquis par des cellules pour accomplir des fonctions cellulaires : l'hélicase réplicative des mitochondries et l'ADN ligase III ATP-dépendante des métazoaires.

a. L'hélicase réplicative DnaB

L'hélicase réplicative DnaB est présente chez toutes les Bactéries, chez de nombreux bactériophages ainsi que chez les Eucaryotes, où l'enzyme est utilisée au cours de la réplication de l'ADN mitochondrial. La phylogénie de la protéine (présentée dans article V) positionne les séquence des bactériophages d'une manière mélangée au sein des bactéries : chaque séquence de bactériophage étant localisée à proximité de la séquence de son hôte (Moreira 2000 ; résultat non montré). Cette phylogénie indique donc que les Bactériophages acquièrent très fréquemment des gènes cellulaires en provenance de leurs hôtes. Néanmoins la position des séquences mitochondriales est étonnante, puisqu'elle ne se positionne pas à proximité des séquences d'Alpha-Protéobactéries mais comme groupe frère du groupe des Bactériophages de type T3/T7. Pour confirmer ce résultat, nous avons effectué une phylogénie avec un nombre restreint de Bactéries afin d'utiliser un nombre plus grand de positions alignées et de pouvoir réaliser une recherche exhaustive en maximum de vraisemblance (temps de calcul beaucoup moins long). La phylogénie est présentée dans l'article V et confirme la relation de parenté entre les séquences mitochondriales et les séquences de Bactériophages de type T3/T7. Cette phylogénie indique que l'Hélicase Réplicative ancestrale des mitochondries d'origine bactérienne a été remplacée par un homologue de type viral. On se trouverait donc dans la même situation que l'ARN polymérase et l'ADN polymérase des mitochondries qui seraient elles aussi originaire d'un Bactériophage apparenté au groupe de T3/T7.

b. Les ADN Ligase ATP-dépendantes.

Il existe deux types non-homologues d'ADN Ligase :
Les ADN Ligases NAD-dépendantes sont uniquement présentes dans le domaine bactérien, la totalité des génomes séquencés contenant au moins un gène codant pour cette enzyme.
Les ADN Ligases ATP-dépendantes sont présentes chez les Eucaryotes (3 "paralogues"), chez les Archéobactéries et chez quelques Bactéries. On trouve aussi ce gène chez des virus Eucaryotes et chez des bactériophages.

Les ADN Ligases ATP-dépendantes se caractérisent par une grande divergence au niveau de la séquence primaire. Lorsque l'on effectue une recherche de type BLAST avec un des gènes Eucaryotes, on rapatrie les trois paralogues Eucaryotes, les gènes d'Archéobactéries, quelques séquences de Bactéries et les séquences des Poxvirus et des Baculovirus. Toutes les autres séquences de Ligase ATP-dépendantes appartenant aux Bactéries, ainsi que certains virus comme T4 ou PBCV1 (virus d'Algue) sortent ensuite avec des valeurs *E* proche ou au delà du seuil de ressemblance significative (valeur *E* égale ou supérieure à 10^{-3}).

La phylogénie des gènes proches des séquences Eucaryotes est indiquée figure 29. Les séquences Eucaryotes forment trois groupes distincts (I, III, IV), les gènes de type I et IV présentant une large répartition phylogénétique, la Ligase III n'étant présente que chez les arthropodes et les mammifères (voir pour revue Martin et Macneill 2002). Dans ce dernier groupe, on observe que les séquences virales se branchent comme groupe frère de leurs hôtes : les Poxvirus avec les mammifères et les virus d'insectes (Baculovirus LdMNPV et Ascovirus) avec *Drosophila melanogaster*. Deux scénarios évolutifs permettent d'expliquer cette observation :

- Soit une duplication chez l'ancêtre des Eucaryotes suivie de pertes sélectives de gènes dans de nombreuses lignées sauf dans celle des arthropodes et des mammifères. Les virus ayant récupéré ces gènes, indépendamment les uns des autres, en provenance de leurs hôtes respectifs.
- Soit les arthropodes et les mammifères ont acquis dans les deux lignées le gène par transfert horizontal en provenance de leurs virus respectifs.

Le deuxième scénario implique moins d'évènements évolutifs, il est donc plus parcimonieux.

Dans cet arbre, les quelques séquences bactériennes apparaissent polyphylétiques et localisées d'une manière robuste au sein des deux groupes d'Archéobactéries. Cette topologie supporte l'idée de transferts horizontaux de ces gènes chez ces Bactéries, en provenance des Archéobactéries. La situation pour les autres séquences bactériennes, très divergentes par rapport aux séquences analysées ci-dessous, est plus délicate à interpréter. Le gène est présent d'une manière ponctuelle chez des groupes phylogénétiquement éloignés (*Bacillus*, *Haemophilus*, *Neisseiria*, *Campylobacter*, *Pseudomonas*). Certaines séquences sont localisées dans des génomes de prophages cryptiques, c'est par exemple le cas pour le gène yoqV chez *Bacillus subtilis* localisé dans le prophage SPβ. Dans les autres cas, l'origine des gènes n'est pas connue. Il pourrait s'agir de transferts horizontaux en provenance des virus et/ou de duplication de gènes suivie de perte sélective dans telle ou telle lignée.

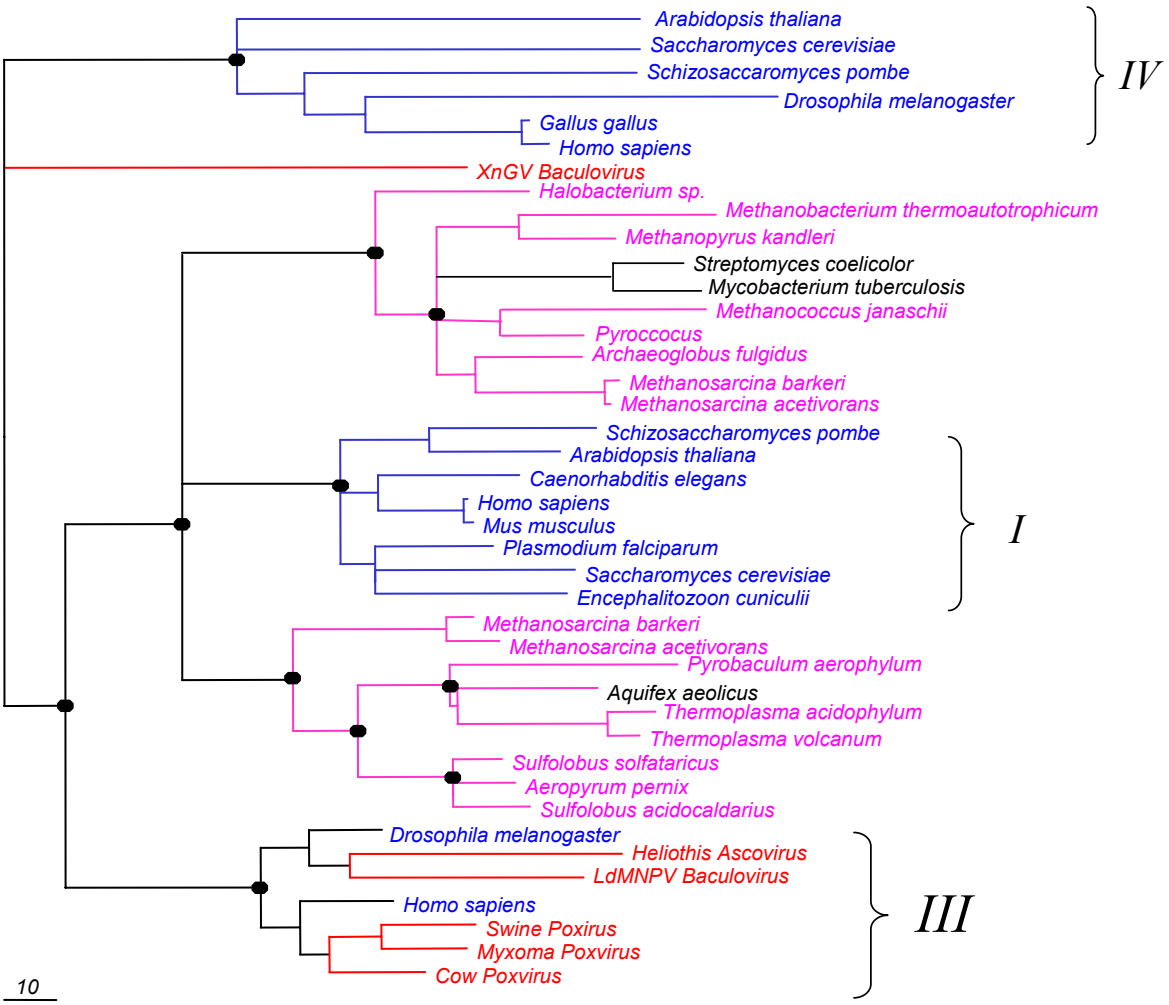

Figure 29: Phylogénie de l'ADN ligase ATP dépendante.

L'arbre est issu d'une recherche rapide en utilisant le programme PROTML (Adachi et Hasegawa 1996) et avec le modèle de substitution en acide aminé JTT-F. Les valeurs de bootstrap sont calculées en utilisant la méthode RELL appliquée aux 1000 meilleurs arbres. Les valeurs supérieures à 95% sont indiquées avec un rond noir. Les Bactéries sont indiquées en noir, les Eucaryotes en bleu, les Archéobactéries en mauve et les virus en rouge. La barre d'échelle représente le nombre de substitutions pour 100 sites par unité de longueur de branche.

c. Autres protéines ayant des homologues viraux

Un certains nombre d'autres protéines cellulaires possèdent des homologues viraux, sans que l'on puisse identifier des transferts horizontaux polarisés des virus vers les cellules. On observe parfois la situation inverse avec de fréquentes acquisitions de gènes cellulaires par des virus, c'est par exemple le cas de la protéine ssb chez les Bactéries (Moreira 2000, résultat non montré).

Parfois les virus codent pour des enzymes qui sont phylogénétiquement très éloignées de leurs hôtes, c'est par exemple le cas des virus Eucaryotes codant le facteur de processivité PCNA. Une situation de ce type est aussi observable pour la Rnase HI du Bactériophage T5. Le cas des Rnases est d'autant plus intéressant qu'il existe une autre enzyme non homologue, la Rnase HII qui est universellement conservée. La Rnase HI est présente chez presque tous les Eucaryotes et toutes les Bactéries. Leipe et collaborateurs (1999) avaient proposé que cette homologie entre gènes Eucaryote et Bactérien pouvait résulter de transfert horizontaux de gène entre ces organismes, via par exemple l'évènement d'endosymbiose mitochondriale.

La phylogénie de cette enzyme présentée figure 30 indique que l'enzyme Eucaryote n'est probablement pas issue de l'endosymbiose d'une mitochondrie, car les séquences Eucaryotes ne sont pas phylogénétiquement proches des α-Protéobactéries. La monophylie des groupes Bactériens et Eucaryotes est statistiquement bien soutenue, l'hypothèse d'un transfert horizontal ancien entre Bactérie et Eucaryote n'est donc pas plus parcimonieuse que l'hypothèse concurrente postulant que le LUCA possédait le gène qui a été perdus dans la lignée des Archéobactéries. On peut constater que le génome de *Vibrio cholerae* contient deux gènes codant une Rnase HI, très éloignés phylogénétiquement l'un de l'autre et dont l'un se branche comme groupe frère du phage T5. Ce dernier gène étant donc probablement d'origine virale et l'autre annoté "VC2234" étant le gène cellulaire ancestral des γ-Protéobactéries. La phylogénie indique aussi que la séquence du Bactériophage T5 se branche à proximité de la base de l'arbre des Bactéries, cette position ne permet donc pas de rejeter l'hypothèse d'une origine virale pour la Rnase HI de l'ensemble des Bactéries. Toutefois, nous avons besoin d'un nombre plus grand de séquences de Virus pour préciser ce résultat.

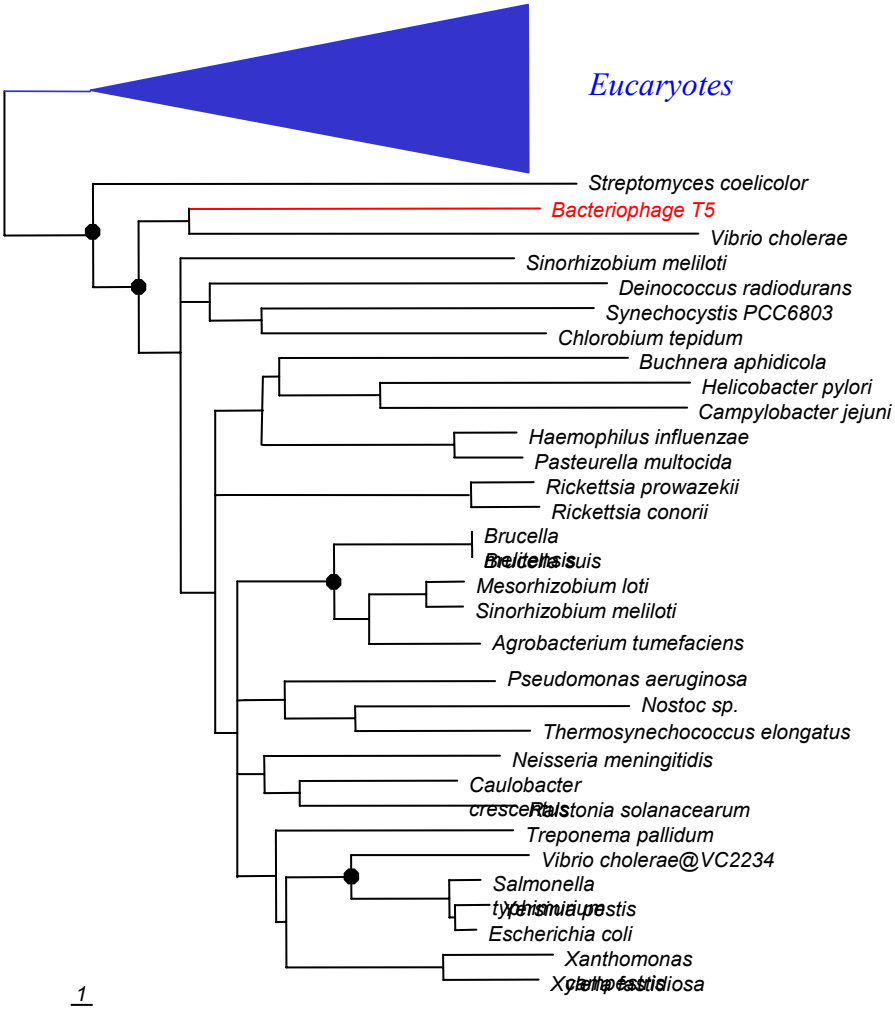

Figure 30 : Phylogénie de la Rnase HI

L'arbre est issu d'une recherche rapide en utilisant le programme PROTML (Adachi et Hasegawa 1996) et avec le modèle de substitution en acide aminé JTT-F. Les valeurs de bootstrap sont calculées en utilisant la méthode RELL appliquée aux 1000 meilleurs arbres. Les valeurs supérieures à 95% sont indiquées avec un rond noir. Les Bactéries sont indiquées en noir, les Eucaryotes en bleu, les virus en rouge. La barre d'échelle représente le nombre de substitutions pour 100 sites par unité de longueur de branche.

CHAPITRE G

Enzymes impliquées dans le métabolisme de l'ADN

Nous allons développer au cours de ce chapitre l'idée selon laquelle les virus ont joué un rôle important au cours de l'histoire évolutive des enzymes impliquées dans le métabolisme terminal des acides nucléiques. Pour illustrer ce concept nous allons discuter en détail de l'histoire évolutive des enzymes suivantes :

- Les Thymidylate synthases et la Dihydrofolate réductase (synthèse du dTMP)
- Les Ribonucléotides réductases (synthèse de dNTP/dNDP)
- La dUTPase et la dCMP déaminase (synthèse du dUMP)

a. La synthèse du dTMP

La Thymidylate Synthase catalyse la formation de dTMP à partir de dUMP. Cette enzyme est indispensable à tout être vivant car les Ribonucléotide réductases ne permettent pas de produire directement du déoxythymidylate. Jusqu'a récemment, on ne connaissait qu'une seule enzyme, codée par les homologues du gène ThyA, capable de catalyser cette réaction. Pourtant aucune thymidylate synthase ThyA n'avait été détectée dans le génome de nombreux microorganismes. En 2002, Hannu Myllykallio et ses collaborateurs ont démontré biochimiquement qu'une autre Thymidylate Synthase, non-homologue à ThyA, existait dans le vivant. Produit du gène ThyX, ce gène a été détecté parce qu'il possédait, à une exception près (voir ci-dessous), une répartition phylogénétique complémentaire au gène ThyA : lorsque ThyA est présent, ThyX est absent et inversement. La répartition phylogénétique des deux gènes est tout a fait atypique puisque des organismes très proches peuvent ne pas partager cette enzyme : la phylogénie de ThyA est visible dans l'article V et celle de ThyX dans l'article III (en fin de chapitre). Cette répartition "en patch" des enzymes est très probablement le résultat de transferts horizontaux de gènes, suivis ou non de remplacement non homologue d'une des copies. Deux cas semblent bien documentés. D'une part, le protiste *Dyctiostellium discoideum* qui est le seul Eucaryote à posséder une Thymidylate synthase ThyX, très proche phylogénétiquement des séquences d'Alpha-Protéobactéries, alors que tous les autres Eucaryotes possèdent un homologue du gène ThyA. D'autre part les Bactéries du genre *Mycobacterium* qui sont les seuls organismes à posséder à la fois ThyX et ThyA ; la séquence du gène ThyX étant très proche dans l'arbre des séquences virales des Mycobactériophage D5 et D29. Cette proximité indiquent très probablement l'acquisition récente du gène viral par les Bactéries du genre *Mycobacterium*. D'autres cas de transferts horizontaux entre virus et cellules peuvent aussi être mis en évidence. Pour le gène ThyX, les séquences des

Archéobactéries du genre *Halobacterium,* localisées en groupe frère du phage d'halophile HF2 (résultat non montré de la phylogénie plus récente), se branchent d'une manière distante des autres Archéobactéries. Cet éloignement phylogénétique est un bon critère pour penser que c'est bien la cellule qui a acquis le gène viral et non l'inverse. Pour le gène ThyA, on peut remarquer que les génomes des Bactéries du genre *Bacillus* possèdent deux gènes codant pour cette enzyme. Ces deux "paralogues" sont phylogénétiquement très éloignés l'un de l'autre et l'un des deux gènes est positionné dans l'arbre en tant que groupe frère du Bactériophages beta-22 qui parasite *Bacillus subtilis.* Ainsi, cette copie cellulaire est aussi vraisemblablement d'origine virale.

Un troisième gène est impliqué dans la synthèse du dTMP. Il s'agit de la Dihydrofolate reductase (DHFR). Biochimiquement, ce gène est intimement associé à l'activité de ThyA. En effet ThyA catalyse la formation du dTMP en produisant du dihydrofolate, ce dernier composant étant réduit en tétrahydrofolate par la DHFR pour être utilisé dans diverses voies métaboliques. L'autre Thymidylate Synthase (ThyX) produit directement du tetrahydrofolates sans avoir besoin de la DHFR. Or l'analyse des génomes complets démontre que la DHFR est parfois aussi présente chez des Bactéries possédant la Thymidilate Synthase ThyX . Quel pourrait être l'origine évolutive de ces gènes codant pour la DHFR ? Deux scénarios étant envisageables :

- soit la DHFR est conservée lors du remplacement de ThyA par ThyX pour des raisons restant à déterminer.
- soit le gène codant pour la DHFR est acquis par transfert horizontal, indépendamment de l'histoire évolutive de la Thymidylate synthase présente dans le génome correspondant.

Pour tester ces hypothèses nous avons effectué la phylogénie de ce gène, présentée dans l'article IV. Dans cette phylogénie, on peut d'abord noter l'influence des plasmides et transposons pour le remplacement homologue de la copie cellulaire de certaines souches d'*E. coli* en relation avec la résistance à un antibiotique : le triméthoprim. Les Bactéries possédant à la fois ThyX et la DHFR sont indiquées en bleu. Pour les espèces du genre *Clostridium,* on peut remarquer que la localisation de ces gènes dans l'arbre est proche d'une séquence plasmidique loin des autres Bactéries Gram+ à bas taux GC (Firmicutes). Il est donc possible que ce gène de *Clostridium* ait été acquis par transfert horizontal en provenance d'un plasmide. Une situation similaire est visible dans l'arbre pour *Rickettsia conorii,* localisée très loin des autres Alpha-Protéobactéries. Il s'agirait probablement là aussi d'un transfert de gène.

Pour les autres taxa possédant ThyX et DHFR on ne peut pas se prononcer car leur placement phylogénétique dans les arbres en général est problématique (cas de *Thermotoga* et des taxons à grandes branches comme les *Chlamydia*). Ainsi, la présence conjointe de ThyX et de la DHFR s'expliquerait plutôt par des transferts de gènes de cette dernière enzyme, par exemple en provenance de virus ou de plasmides.

Pris ensemble, tout ses résultats indiquent que l'histoire évolutive des Thymidylate synthases est marquée par de nombreux événements de transferts de gènes, entre particulier entre virus et cellules. Ces transferts de gène peuvent être suivis du remplacement non-homologue de la copie initialement présente, ce qui a pour conséquence une répartition phylogénétique de chaque gènes particulièrement complexes.

b. La Synthèse des dNDP/dNTP

Il existe trois familles de Ribonucléotide réductases présentant toutes les trois des similarités structurales et mécaniques, mais seule la sous-unité catalytique des familles I et II présentant de fortes similarités de séquences primaires. La répartition des deux familles homologues (I/II et III) dans les trois domaines ne suit pas la répartition habituellement observée pour les protéines informationnelles. La famille I/II est universellement conservée, tandis que la famille III est présente d'une manière sporadique chez les Procaryotes. La phylogénie globable de la famille I/II issue de la fusion des deux sous-unités est présentée en annexe dans le chapitre de livre. Pour préciser les relations phylogénétiques entre bactériophages et cellules nous avons effectué les phylogénies avec un nombre moins élevé de taxons (plus grand nombre de positions alignées utilisable). La phylogénie de la grande sous-unité est présentée dans l'article V (les phylogénies de la petite sous-unité et de la fusion des deux sous-unités donnent des résultats identiques, résultats non montrés).

Cette phylogénie montre un cas intéressant de transfert horizontal polarisé du virus vers son hôte dans le cas des Protéobactéries. En effet les Bactéries appartenant aux espèces *Salmonella typhimurium*, *Escherichia coli* et *Neisseria meningitidis* possèdent deux copies du gènes dans leurs génomes : un gène de type "I a" et un gène de type "I b". Ces deux "paralogues" sont phylogénétiquement très distants. Pour le gène "I a", le phage T4 (ainsi que le phage T5 avec avec une phylogénie plus récente, résultat non montré) se place comme groupe frère de ces Bactéries. Ces deux observations indiquent que le gène Ia de ces Protéobactéries est probablement d'origine virale. Il est possible que les autres Gamma

Protéobactéries ne possédant que la copie "virale" de type I b (*Vibrio*, *Haemophilus*, *Pasteurella*) auraient perdu la copie d'origine "cellulaire" I a.

c. Formation de dUMP

Pour fabriquer du dUMP les cellules ont deux possibilités : soit à partir du dUTP (réaction catalysée par une dUTPase) soit à partir de dCMP (réaction catalysée par une dCMP déaminase). Alternativement, certaines Archéobactéries produisent du dUMP a partir du dCTP grace à une dCTPdeaminase/dUTPase bi-fonctionnelle (Blornberg et la. 2003). Le dUTP étant un composé toxique pour les cellules (intégré par erreur dans l'ADN à la place du dTTP il provoque des cassures de brin) cet enzyme bi-fonctionnelle a l'avantage de transformer directement en dUMP le dUTP produit par la fonction dCTP déaminase.

Chez les Eucaryotes et certains microorganismes comme *Bacillus subtilis* (groupe des Firmicutes, anciennement "gram+ à bas taux de G+C"), le dUMP est directement produit à partir de dCMP sans passer par l'étape intermédiaire de production de dUTP. L'enzyme qui catalyse cette réaction, la dCMP deaminase, est présente chez la plupart des Eukaryotes et des EuryArchéobactéries, ainsi que chez toutes les Bactéries du groupe des Firmicutes. Quelques séquences virales très divergentes sont aussi disponibles. La phylogénie de cette enzyme de petite taille est particulièrement mal résolue et n'est pas véritablement exploitable (résultat non montré).

La dUTPase est une enzyme ayant une très large répartition phylogénétique puisqu'elle est présente chez pratiquement tout les Eucaryotes et toutes les Bactéries. On la retrouve aussi chez de nombreux virus des trois domaines. Il est intéressant de noter que l'enzyme est absente chez de nombreuses Bactéries appartenant aux groupes des Firmicutes : sur 23 génomes séquencés on ne retrouve l'enzyme que dans 7 génomes. Nous avons effectué la phylogénie de l'enzyme chez les Bactéries en racinant l'arbre avec les séquences Eucaryotes. L'arbre est présenté figure 31 . Intéressons-nous aux Firmicutes (indiqués avec une accolade). On constate que les séquences de ces taxons se branchent la plupart du temps en groupe frère des séquences des virus les parasitant. De plus plusieurs gènes cellulaires sont localisés dans des prophages comme celui de *Staphyloccocus aureus* ou un des deux gènes de *Bacillus subtilis*. Ces observations indiquent très probablement que les gènes de dUTPase chez les Firmicutes sont d'origine virale, acquis récemment et indépendamment les uns des autres. Enfin on notera que le deuxième gène de *Bacillus subtilis* qui n'est pas codé par un prophage est voisin du gène codant pour la Thymidylate synthase ThyA précédemment identifié comme

ayant une origine virale. Ainsi il est probable que certaines Bactéries du genre des Firmicutes ont recruté plusieurs fois indépendamment une fonction d'origine virale afin de disposer d'une seconde voie de formation du dUMP, utilisant non pas le dCMP comme substrat, mais le dUTP.

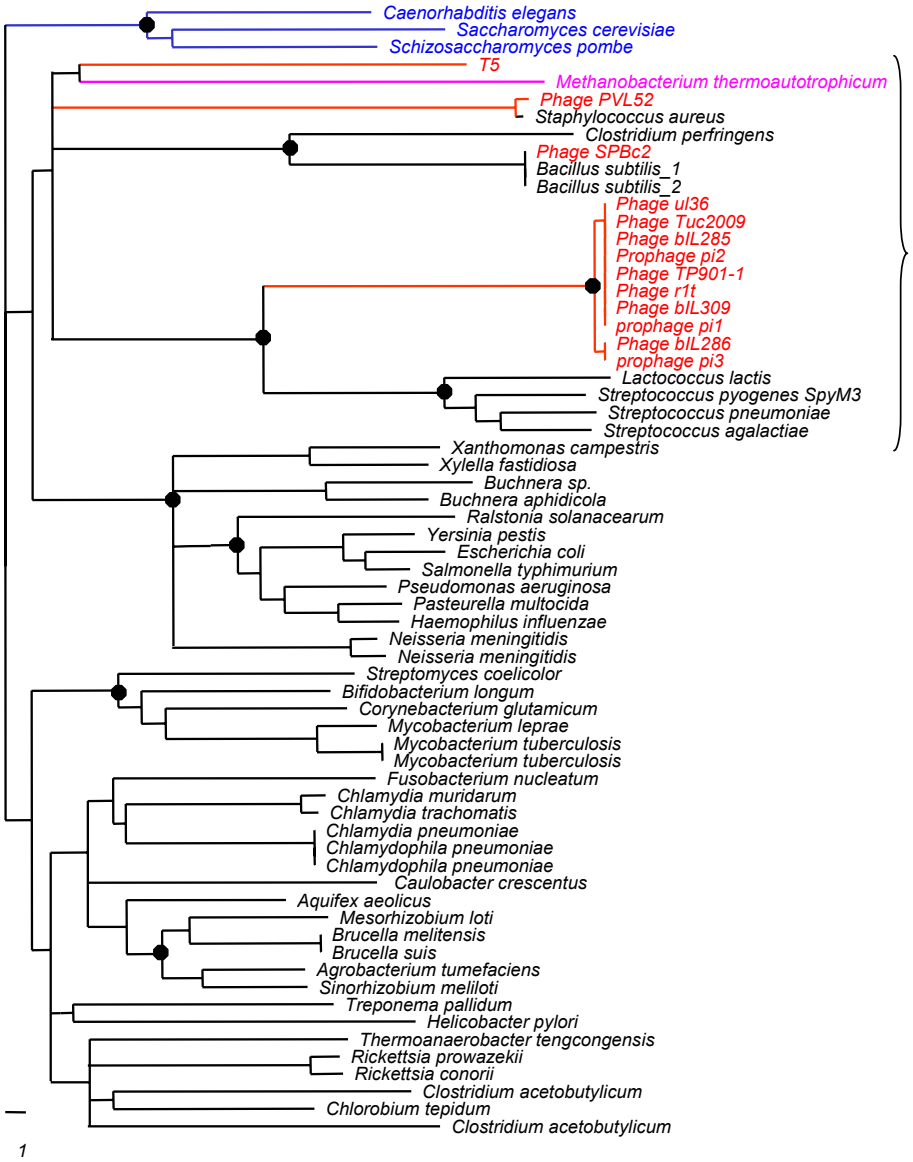

Figure 31 : phylogénie de la dUTPase

L'arbre est issu d'une recherche rapide en utilisant le programme PROTML (Adachi et Hasegawa 1996) et avec le modèle de substitution en acide aminé JTT-F. Les valeurs de bootstrap sont calculées en utilisant la méthode RELL appliquée aux 1000 meilleurs arbres. Les valeurs supérieures à 95% sont indiquées avec un rond noir. Les Bactéries sont indiquées en noir, les Eucaryotes en bleu, les Archéobactéries en mauve et les virus en rouge. La barre d'échelle représente le nombre de substitutions pour 100 sites par unité de longueur de branche.

d. Conclusion

De nombreuses enzymes impliquées dans les étapes terminales de la biosynthèse des acides nucléiques existent sous plusieurs formes d'analogues fonctionnels, sans relation évolutive entre eux. La répartition phylogénétique de ces enzymes est souvent très complexe et implique de nombreux évènements de transfert horizontaux suivi, ou non, de remplacement homologue et non homologue de ces gènes. Nous avons potentiellement détecté plusieurs gènes comme étant d'origine virale. Il s'agit surtout de gènes conservés à petite échelle évolutive parce qu'il est difficile de démontrer l'existence et la polarité de transferts de gène anciens. Néanmoins ces résultats ne permettent pas d'affirmer que les virus sont à la source de la plupart des gènes cellulaires impliqués dans la transition d'un monde ARN vers un monde ADN.

Article III

An alternative Flavin-Dependent Mechanism for Thymidylate synthesis

Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U.

Science (2002) 297 : 105-107.

Article IV

Life without dihydrofolate reductase FolA
Myllykallio H, Leduc D, Filee J, Liebl U.

TIM (2003) 11 : 220-223.

Article V

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

Filee J, Forterre P, Laurent J.

Res. Microbio. (2003) 124 : 237-243

Discussion et Perspectives

A la différence de la machinerie de transcription et de traduction, l'appareil de réplication de l'ADN est très différent entre les Bactéries, d'une part et les Archéobactéries et les Eucaryotes d'autre part. Cette observation intrigue depuis longtemps un certain nombre de chercheurs qui ont proposé différentes hypothèses pour expliquer ces observations. Le séquençage d'un nombre croissant de génomes cellulaires et viraux a permis ces dernières années d'apporter un éclairage nouveau à se sujet. L'objectif de ce travail était d'analyser d'un point de vue phylogénétique les différentes enzymes impliquées dans la réplication et le métabolisme terminal de l'ADN, et en particulier d'évaluer la contribution des virus concernant l'évolution de ces gènes.

a. La répartition phylogénétique des enzymes de la réplication et du métabolisme de l'ADN

Leipe et collaborateurs (1999) avaient publié le détail de la répartition de chaque enzyme impliquée dans la réplication de l'ADN. Ce travail établissait clairement le fossé existant entre les enzymes des Bactéries, d'une part et celle des Eucaryotes/Archéobactéries, d'autre part. Pourtant, en quelques années, le séquençage d'un grand nombre de génomes montrent que cette répartition n'est pas aussi simple. En effet beaucoup d'enzymes semblent partagées entre de nombreux groupes de Bactéries et d'Eucaryotes sans que l'on puisse vraiment soupçonner un transfert horizontal ancien entre ces groupes (par exemple à l'occasion de l'endosymbiose de la mitochondrie). Il s'agit de l'ADN polymérase Y, des ADN Topoisomérases Ib et Ia et de la Rnase HI. Leur présence dans quelques espèces d'Archéobactéries résultent très probablement de transferts horizontaux récents de gènes en provenance des Bactéries. Ajoutons à cela le fait que d'autres enzymes de la réplication semblent bien orthologues entre les trois domaines (RFC, Rnase HII et ADN Ligase ATP-dépendante). Pour les enzymes impliquées dans le métabolisme terminal des acides nucléiques, aucune protéine ne suit la répartition attendue selon Leipe et collaborateurs entre les trois domaines, à l'exception de la dCMP déaminase.

A la lumière de ces résultats, l'hypothèse de Leipe (1999) qui postulait pour une double invention de l'ADN, une fois dans la lignée des Bactéries et une fois dans la lignée des Archéobactéries et des Eucaryotes apparaît peu probable. Toutefois, la plupart de ces enzymes existent généralement sous deux formes non-homologues, une double (ou multiple) invention indépendante de l'ADN reste donc tout à fait envisageable.

Ces résultats tendent à montrer que l'histoire évolutive des composants de la réplication et du métabolisme de l'ADN est très complexe et semble marquée par de nombreux évènements de

remplacement non-homologue d'une enzyme par une autre (Edgel et Doolitle 1997). Effectivement nous avons détecté plusieurs cas de remplacement non homologue, le plus illustratif étant observé avec les Thymidylates synthases. D'autre cas semblent aussi assez probables, citons les Topoisomérases II chez les Archéobactéries du groupe des Thermoplasmatales, les Topoisomérases I chez certains groupes de Bactéries, ainsi que certaines enzymes de la réplication de l'ADN mitochondrial (voir paragraphe c). Ce dernier exemple est particulièrement intéressant puisque la source des enzymes impliquées dans le remplacement non-homologue se trouve être un virus.

b. L'influence des virus

Patrick Forterre (1999) et Luis Villarreal (2000) proposent que de nombreuses enzymes de la réplication ont été inventées par les virus, et que ces enzymes ont remplacé des non-homologues cellulaires. Cette hypothèse est séduisante parce qu'elle permet d'expliquer la grande variabilité de la répartition phylogénétique des différentes enzymes entre les trois domaines du vivant. Résiste t'elle à l'épreuve des phylogénies ?

Nous avons identifié un certain nombre de gènes potentiellement d'origine virale occupant des fonctions cellulaires, la figure 32 résume ces observations. Il est frappant de constater que la presque quasi-totalité des transferts identifiés concerne des Bactéries. Pour les Archéobactéries le très faible échantillonnage de génomes viraux pourrait expliquer l'impossibilité de détecter la présence de gène cellulaire d'origine virale dans leurs génomes. Mais pour les Eucaryotes, et singulièrement les métazoaires, cette explication ne tient plus. Les Bactéries seraient elles quantitativement plus affectées par les transferts horizontaux en provenance de virus que les Eucaryotes ? Cette observation est elle liée au fait que les virus Eucaryotes adoptent beaucoup moins souvent des stratégies lysogènes que les bactériophages ? Toutefois, même chez les Bactéries, le problème de l'échantillonnage en séquences virales est crucial. Si l'on excepte le cas des organelles, la presque totalité des transferts identifiés concerne les deux même groupes d'organismes : les γ-Protéobactéries et les Firmicutes ("Gram positive à bas taux de G+C"). Ces deux groupes sont largement étudiés, incluent deux organismes "modèles" en biologie moléculaire (respectivement *Escherichia coli* et *Bacillus subtilis*). Un grand nombres de génomes cellulaires et surtout viraux infectant des organismes de ce groupe est disponible. Il semble assez clair que notre connaissance très partielle de la biodiversité des virus limite l'identification des gènes

cellulaires d'origine virale dans les génomes de très nombreux groupes Bactériens. On minimise donc l'influence globale des virus de ce point de vue.

Enzyme	Organismes	Remarques
Ribonucléotide Reductase (2 ssu.)	Gamma Protéobacteries Beta Protéobacteries	« Paralogue » Ia
Thymidylate Synthase ThyA	Gamma Proteobacteries *Bacillus*	
Thymidylate Synthase ThyX	Actinobacteries *Halobacterium*	
Dihydrofolate reductase	Escherichia Clostridium	Résistance aux antibiotiques
ADN polymérase B	Gamma Proteobacteries Mitochondries de Plantes *Halobacterium* Eucaryotes	« Paralogue » Delta
ADN polymérase A	Mitochondries (hors Plante)	
ADN ligase ATP-dépendante	Métazoaires	« Paralogue » III
Rnase HI	*Vibrio*	
Hélicase Réplicative DnaB	Mitochondries	
ARN polymérase T3/T7	Mitochondries Chloroplastes	
dUTPase	Firmicutes	

Figure 32 : Gènes cellulaires potentiellement d'origine virale.

Le nom du gène est indiqué, les taxa possédant le gène viral sont présentés en noir pour les Bactéries et les organelles, en bleu pour les Eucaryotes et en mauve pour les Archéobactéries.

On peut aussi constater que la plupart des transferts horizontaux de gène polarisés des virus vers les cellules qui ont été détectés sont des évènements à petite échelle évolutive, à l'exception notable de l'appareil de réplication de la mitochondrie. Il est possible que l'acquisition de gènes viraux par les cellules soit un processus fréquent, mais, étant donné que ces évènements ne donnent généralement pas d'avantage sélectif majeur, la plupart de ces gènes sont secondairement perdus. Si le gène viral reste conservé à grande échelle évolutive, sa détection pose ensuite des problèmes méthodologiques. En effet, de très nombreux gènes viraux sont phylogénétiquement très éloignés de ceux de leurs hôtes, parfois ceux ci se branchent autour de la base du groupe considéré. Plusieurs hypothèses permettent d'expliquer ce cas de figure :

1) Absence de transfert horizontal du gène entre le virus et la cellule.
2) Acquisition du gène cellulaire par le virus suivi d'une forte accélération de la vitesse d'évolution du gène. Le gène viral se place incorrectement dans les phylogénies sous l'effet d'attraction des longues branches.
3) Acquisition ancienne du gène viral par les cellules, antérieurement à la divergence de nombreux taxa.

Seule l'utilisation d'un critère *a priori* peut nous permettre de polariser la direction du transfert. Or il n'existe malheureusement pas de phylogénie universelle "de référence". On doit donc composer avec les quelques éléments bien établis en biologie de l'évolution comme par exemple l'origine bactérienne des mitochondries dont nous pensons qu'au moins une partie de l'appareil de réplication pourrait être d'origine virale.

c. L'appareil de réplication de l'ADN des mitochondries est il, en partie, d'origine virale ?

Le mécanisme général de la réplication de l'ADN de la mitochondrie (ADNmt) n'est pas connu avec certitude, toutefois il semblerait que le mécanisme soit continu (synthèse couplée des deux brins) au moins chez les mitochondries des mammifères (Yang et al. 2002). Les enzymes connues pour être impliquées dans la réplication de l'ADNmt sont généralement codées dans le noyau et les enzymes importées dans l'organelle. Le mécanisme d'initiation de l'ADN mitochondrial n'est pas élucidé, soit la Réplicase (ADN polymérase γ de la famille A) possède une activité Primase, soit c'est l'ARN polymérase ADN dépendante impliquée dans la transcription des quelques gènes mitochondriaux qui réalise la synthèse des amorces (Yang et al. 2002). Cette ARN polymérase n'est pas homologue de l'ARN polymérase des bactéries, mais uniquement d'une ARN polymérase de virus et de plasmides (Gray et Lang 1998). La

seule exception connue concerne les mitochondries du protiste *Reclinomonas americana* qui ont conservé une ARN polymérase bactérienne (Lang et al. 1997). L'hypothèse proposée postule que l'ARN polymérase bactérienne ancestrale des mitochondries a été remplacée par une ARN polymérase de virus appartenant aux phages du groupe de T3/T7.
Nous avons potentiellement identifié deux autres enzymes impliquées dans la réplication de l'ADN mitochondrial qui pourrait être originaires d'un virus : l'Hélicase Réplicative DnaB et l'ADN polymérase famille A. Dans les deux cas, l'enzyme mitochondriale est phylogénétiquement proche des homologues codés par les phages du groupe de T3/T7 et très distante ou non-homologue de l'enzyme des α-protéobactéries. On peut donc penser à un remplacement d'au moins trois gènes cellulaires par des contreparties virales originaire d'un phage du groupe de T3/T7. Le transfert horizontal de ces trois gènes à la fois, en provenance d'un phage du groupe T3/T7, est documenté dans le génome de *Pseudomonas putida* KT40 qui contient un prophage complet intégré dans son chromosome. Toutefois, on ne peut pas écarter complètement l'idée que la proximité des séquences virales avec les séquences mitochondriales résulte d'un artefact d'attraction des longues branches entre ces enzymes très divergentes comparées aux protéines bactériennes. Mais cet artefact ne permet pas d'expliquer pourquoi ces gènes mitochondriaux sont à chaque fois phylogénétiquement proche de ceux des phages du groupe de T3/T7.

Chez les mitochondries de plantes, on ne retrouve pas d'ADN polymérase de la famille A. La mitochondrie des plantes possède pour seule ADN polymérase mitochondriale une ADN polymérase de la famille B dite "protein primed", portée par un plasmide linéaire. Il est donc probable que cette ADN polymérase plasmidique (famille B) ait remplacé l'ADN polymérase de type T3/T7 (famille A) qui elle même remplaçait l'ADN polymérase de l' α-Protéobactérie ancestrale qui était à la source de la mitochondrie (famille C).

Le remplacement des composants bactériens par des composants viraux peut s'être produit soit avant l'endosymbiose chez l'α-Protéobactérie ancestrale, soit après l'endosymbiose. A notre connaissance personne n'a jamais isolé de virus parasitant des mitochondries, et il ne serait pas étonnant que les α-Protéobactéries à l'origine de la mitochondrie utilisaient un système de réplication de l'ADN légèrement divergent par rapport à celui utilisé par les α-Protéobactéries actuelles. Alternativement, l'α-Protéobactérie à l'origine de la mitochondrie pourrait, comme dans le cas de *Pseudomonas putida* KT40, avoir intégré un prophage

appartenant au groupe de T3/T7. Certains gènes viraux auraient été conservés et exportés dans le noyau, les contre parties "cellulaires" auraient été perdues. Cette hypothèse a l'avantage de permettre d'expliquer simplement le cas du protiste *Reclinomonas americana* qui possède une ARN polymérase bactérienne et non virale : c'est l'ARN polymérase de type bactérienne qui a été retenue et non le type viral a l'inverse de la plupart des autres Eucaryotes. Le séquençage de génomes complets de différents protistes permettra de savoir si ces organismes possédent une Hélicase Réplicative DnaB et une ADN polymérase de la famille A de type viral ou plutôt bactérienne.

Il est intéressant de noter que d'autres composants de la réplication de l'ADNmt ne sont pas d'origine bactérienne. C'est le cas de la RnaseHI (Cerritelli et al. 2003) et de la Topoisomérase III α (Wang et al. 2002) qui sont des composants "eucaryotes" qui co-localisent à la fois dans la mitochondrie et le noyau. Enfin d'autres composants sont homologues de composants bactériens (et de bactériophages) comme la protéine ssb (De Pamphilis 1996). La machinerie de la réplication de l'ADN de la mitochondrie apparaît donc comme complètement chimérique entre des éléments bactériens, eucaryotes et viraux. Cela illustre assez bien l'apparente "plasticité" de l'appareil de réplication de l'ADN, alors que l'on s'attendrait plutôt à trouver un système stable et très conservé car impliqué dans de grands complexes macro-moléculaires constitués de multiples interactions entre protéines. Pour autant, étant donné que l'histoire évolutive de la mitochondrie est un cas de figure très particulier dans le vivant, l'évolution de son appareil de réplication ne peut probablement pas être extrapolé à tout le vivant.

d. Les virus ont ils inventé l'ADN ?

Un certain nombre de gènes cellulaires impliqués dans la réplication et le métabolisme de l'ADN semble avoir une origine virale. La question qui se pose est la suivante : est-ce que ces gènes ont été effectivement "inventés" par les virus ou s'agit il d'un simple transport de ces gènes de cellule à cellule.

On peut en effet concevoir les virus comme de simples transporteurs de gènes d'une espèce cellulaire à une autre : ce cas de figure est bien documenté pour les gènes de résistance aux antibiotiques ou localisé dans les "îlots de pathogénicité". Néanmoins, de très nombreux gènes viraux impliqués dans la réplication sont très divergents au niveau de la séquence primaire par rapport à des gènes cellulaires. Les virus ne peuvent donc pas être conçus comme effectuant seulement un "auto-stop" génique d'une espèce cellulaire à une autre. On peut par

exemple penser que les virus acquièrent des gènes cellulaires ("morons"), ces gènes divergent et peuvent connaître des changements de fonctions par relaxation de la pression de sélection (le virus peut continuer à utiliser les homologues cellulaires "originaux" de ce gène). Secondairement, les cellules peuvent récupérer ce gène viral et bénéficier d'une nouvelle fonction.

Certains génomes viraux codent aussi pour des enzymes qui n'ont aucun homologue cellulaire connu (par exemple l'ADN polymérase de la famille E) et qui pourraient donc être considérées comme ayant été "inventées" en tant que telles par les virus. Il n'est donc pas impossible qu'une partie de ces gènes inventés par les virus puisse avoir été recrutés secondairement par les cellules. Néanmoins, compte tenu de nos connaissances très partielles de la biodiversité des organismes cellulaires, on ne peut pas être certains que ces enzymes virales n'ont effectivement pas d'homologues cellulaires. Et inversément, le fait qu'une enzyme cellulaire ne possède pas d'homologues viraux connus n'indiquent pas que les virus n'ont joué aucun rôle dans l'évolution de cette protéine.

e. Perspectives

Les virus ont pendant des décenies été considérés comme de simples éléments égoïstes des génomes, produits récents de l'autonomisation de morceaux d'ADN cellulaire. En montrant que les virus étaient des éléments anciens, remontant sans doute, pour certains d'entre eux, antérieurement à la divergence des trois domaines du vivant, le séquençage d'un grand nombre de génomes viraux a permis de remettre en cause cette vision. Ces découvertes ont ouvert de nombreuses portes concernant la reconstitution de l'histoire évolutive de ces éléments et l'impact des virus sur l'évolution de leurs hôtes. Pour autant de nombreuses questions restent en suspens, principalement parce que notre connaissance de la biodiversité des virus est très parcellaire. Le cas des bactériophages du groupe T3/T7 est tout à fait illustratif. Ce groupe de phages demeure l'élément dominant dans les communautés virales marines (Breitbart et al. 2002), il comprend des phages infectant des hôtes bactériens très éloignés phylogénétiquement et occupant des niches écologiques très diverses (depuis des Entérobactéries du tube digestif humain jusqu'à des Cyanobactéries marines). Nous avons montré dans ce travail que certaines enzymes mitochondriales pourraient être originaires d'un virus appartenant à ce groupe. La capacité des virus à s'intégrer dans les génomes cellulaires comme chez Pseudomonas putida KT40 fait de ces éléments des acteurs potentiellement importants concernant l'évolution de leurs hôtes.

Nous avons aussi largement abordé au cours de ce travail l'évolution de l'appareil de réplication de l'ADN. En 1999, Leipe et collaborateurs avaient mis en évidence les profondes différences entre l'appareil de réplication des Bactéries d'une part, et des Archéobactéries et des Eucaryotes d'autre part En quelques années, le séquencage d'un nombre beaucoup plus grand de génomes remet en cause cette vision des choses et indique que la répartition phylogénétique de tel ou tel gène ne suit souvent pas la distribution attendue. On connaît pour l'instant surtout des génomes d'organismes pathogènes, mais, dans les années à venir, le séquencage d'un plus grand nombre de génomes complets de protistes et de procaryotes de "l'environnement", nous permettra surement de préciser les différentes observations effectuées dans ce travail.

Enfin, nous avons maintes raisons d'être optimiste concernant les problématiques évolutives des virus. Le séquencage "en routine" de collections de phages et de virus de plus en plus éloignés phylogénétiquement ainsi que le développement des techniques d'analyse génomique de communautés virales non-cultivées ouvrent, dès a présent, un fantastique champ exploratoire à ceux qui voudront s'y intéresser.

Annexe A : liste des principaux programmes utilisés

Dans cette annexe, le lecteur pourra trouver les principaux programmes informatiques utilisés ainsi qu'une brève description de leurs usages :

- BLAST et PSI-BLAST (Altschul et al. 1997) :
Ensemble de logiciels de recherche de séquences homologues par similitude de séquences dans des banques de données locales ou en ligne.
Nous avons essentiellement utilisé les programmes installés sur site Internet du NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) et sur le serveur local de notre laboratoire (http://www-archbac.u-psud.fr/Projects/Pab_r/Blast2_SeqQuer.html).

- alibaba (Philippe Lopez, communication personnelle) :
Ce programme installé sur machine UNIX Sun Solaris© permet a partir du fichier de sortie du programme BLAST de rapatrier les séquences choisies aux format GenBank, PIR et Swiss-Prot et les concaténer les unes aux autres au format FASTA "plat".

- orf_retriever (Yvan Zivanovic, non publié) :
Ce programme en ligne (http://www-archbac.u-psud.fr/Genomap/orf_retriever.html) permet de rapatrier automatiquement les séquences choisies grâce à leurs code d'identification (format GenBank).
Le serveur BLAST du laboratoire a donc été modifié pour automatiser directement la procédure depuis le fichier de sortie du programme BLAST. Le programme orf_retriever a été légèrement modifié en ajoutant une option permettant de concaténer toutes les séquences sous un format FASTA plat utilisable par le logiciel d'alignement de séquence MUST (option no_def).

- Genomapper (Yvan Zivanovic, non publié) :
Ce logiciel en ligne (http://www-archbac.u-psud.fr/genomap/GenomapBrowser.html) permet de visualiser le contexte génomique d'un gène et de comparer avec le contexte génomique de ses homologues chez d'autres espèces.

- fus2ali :
Ce petit utilitaire fonctionnant sous environnement UNIX Mac OSX© permet de fusionner les différentes sous-unités d'un gène appartenant à une même espèce. Le programme reconnaît les

2 séquences ayant le même identifiant dans 2 fichiers différents au format FASTA et il les concatène dans un fichier de sortie au format FASTA.

- CLUSTALW (Thompson et al. 1994) :
Ce programme installé sur machine UNIX Mac OS X© permet d'établir l'alignement multiple de séquences nucléiques ou protéiques.
(http://evolution.genetics.washington.edu/phylip.html)

- MUST (Philippe 1993) :
Cet ensemble de programmes DOS installés sous PC-Pentium© permettent d'aligner manuellement des séquences en format FASTA, de calculer des arbres avec la méthode du Neighbor-Joining et convertir le fichier aligné en plusieurs formats reconnus par différents logiciels de phylogénie (NEXUS, PHYLIP).

- PROTML (Adachi et Hasegawa 1996) :
Logiciel de reconstruction de phylogénies à partir de séquences protéiques utilisant la méthode du maximum de vraisemblance (inclus dans le package MOLPHY : ftp.ism.ac.jp:/pub/ISLMB/MOLPHY). Ce programme a été installé sur machine UNIX Sun Solaris© et Mac OS X©

- PUZZLE (Strimmer et Von Haeseler 1996) :
Ce programme de reconstruction phylogénétique par maximum de vraisemblance installé sur machine UNIX MacOS X permet en particulier de tenir compte des différences de vitesses d'évolution entre les sites grâce à une loi gamma. Il permet aussi de tester la topologie des arbres obtenus avec des statistiques de type Kishino-Hasegawa (Kishino et Hasegawa 1989).
(http://www.tree-puzzle.de/).

- STRING (von Mering et al. 2003) :
Ce logiciel en ligne (http://www.bork.embl-heidelberg.de/STRING/) permet d'évaluer la répartition phylogénétique d'un gène chez un grand nombre d'espèces et de déterminer si la présence du gène est statistiquement liée à la présence d'un autre gène parmi tous ces génomes.

- KEGG :

Ce site Internet compile toute les données relatives aux différentes voies métaboliques présentes dans le vivant. Il permet en particulier d'évaluer *in silico* la présence de telles ou telles voies métaboliques ou de telles ou telles enzymes et ce, pour un grand nombres d'espèces différentes (http://www.genome.ad.jp/kegg/pathway/map/map01140.html).

Annexe B Chapitre de livre :

Origin and Evolution of DNA and DNA Replication Machineries
Patrick Forterre, Jonathan Filée et Hannu Myllykallio
In The Genetic Code and the Origin of life, 2003, ed. Luis Ribas de Poupana.

Références bibliographiques

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

Alba MM, Das R, Orengo CA, Kellam P (2001) Genomewide function conservation and phylogeny in the Herpesviridae. Genome Res. 11:43-54

Allison GE, Angeles DC, Huan P, Verma NK (2002) Morphology of temperate bacteriophage SfV and characterisation of the DNA packaging and capsid genes: the structural genes evolved from two different phage families. Virology 308:114-27.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-402

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 59:143-69

Atkins JF (1993) Contempory RNA genome. In the RNA world. Cold Spring Harbor Laboratory Press, NY.

Avery OT, Macleod CM, McCarthy M (1944) Studies on the chemical nature of the substance inducing transformation of pneumoccocal type. J Exp Med 79:137-157

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet.11:442-4

Bamford DH, Burnett RM, Stuart DI (2002) Evolution of viral structure. Theor Popul Biol. 61:461-70

Bandea CI (1983) A new theory on the origin and the nature of viruses. J Theor Biol. 105:591-602

Baldauf SL, Palmer JD, Doolittle WF (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc Natl Acad Sci USA 93:7749-54

Baldo A, McClure M (1999) Evolution and horizontal transfer of dUTPase-encoding genes in viruses and their hosts. J Virol. 73:7710-7721

Balter M (2000 Virology. Evolution on life's fringes. Science 5486:1866-7

Bargonetti J, Reynisdottir I, Friedman PN, Prives C (1992) Site-specific binding of wild-type p53 to cellular DNA is inhibited by SV40 T antigen and mutant p53. Genes Dev. 6:1886-98.

Barns SM, Delwiche CF, Palmer JD, Pace NR (1996) Perspectives on archaeal diversity, thermophily and monophily from environemental rRNA sequences. Proc Natl Acad Sci USA 93:9188-9193

Bell PJ (2001) Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? J Mol Evol. 53:251-6.

Benson SD, Bamford JK, Bamford DH, Burnett RM (1999) Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. Cell. 98:825-33

Berg DE, Berg CM, Sasakawa C (1984) Bacterial transposon Tn5: evolutionary inferences. Mol Biol Evol. 5:411-22

Bergh O, Borsheim KY, Bratbak G, Heldal M (1989) High abundance of viruses found in aquatic environments. Nature. 6233:467-8

Bjornberg O, Neuhard J, Nyman PO (2003) A bifunctional dCTP deaminase-dUTP nucleotidohydrolase from the hyperthermophilic archaeon Methanocaldococcus jannaschii.J Biol Chem. 278:20667-72

Blaisdell BE, Campbell AM, Karlin S (1996) Similarities and dissimilarities of phage genomes.Proc Natl Acad Sci USA 93:5854-9

Boltner D, MacMahon C, Pembroke JT, Strike P, Osborn AM (2002) R391: a conjugative integrating mosaic comprised of phage, plasmid, and transposon elements. J Bacteriol. 184:5158-69

Bowen NJ, Jordan IK Transposable elements and the evolution of eukaryotic complexity. Curr Issues Mol Biol. 4:65-76

Brassard S, Paquet H, Roy PH (1995) A transposon-like sequence adjacent to the AccI restriction-modification operon. Gene 157:69-72

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities Proc Natl Acad Sci USA 2002 99:14250-5

Brochier C, Philippe H, Moreira D (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends Genet.16:529-33

Brochier C, Philippe H. (2002) Phylogeny: a non-hyperthermophilic ancestor for bacteria. Nature (6886):244.

Brown JR, Doolittle WF (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci USA 92:2441-5

Brussow H (2001) Phages of dairy bacteria. Annu Rev Microbiol. 55:283-303

Cambillau C, Claverie JM (2000) Structural and genomic correlates of hyperthermostability. J Biol Chem.275:32383-6

Capy P, Vitalis R, Langin T, Higuet D, Bazin C (1996) Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? J Mol Evol. 42:359-68

Castresana J, Moreira D (1999) Respiratory chains in the last common ancestor of living organisms. J Mol Evol. 49:453-60

Castresana J (2001) Comparative genomics and bioenergetics. Biochim Biophys Acta. 1506:147-62

Cerritelli SM, Frolova EG, Feng C, Grinberg A, Love PE, Crouch RJ (2003) Failure to produce mitochondrial DNA results in embryonic lethality in Rnaseh1 null mice. Mol Cell. 3:807-15

Collins RF, Gellatly DL, Sehgal OP, Abouhaidar MG (1998) Self-cleaving circular RNA associated with rice yellow mottle virus is the smallest viroid-like RNA. Virology 241:269-75

De Pamphilis M (1996) DNA replication in Eukaryotic cells. Cold Spring Hrabor Laboratory Press. NY

Daubin V, Perriere G (2003) G+C3 structuring along the genome: a common feature in prokaryotes. Mol Biol Evol. 20:471-83

Di Giulio M (2000) The universal ancestor lived in a thermophilic or hyperthermophilic environment. J Theor Biol. 203:203-13

Di Giulio M (2003) The universal ancestor was a thermophile or a hyperthermophile: tests and further evidence. J Theor Biol. 221:425-36.

Diez B, Pedros-Alio C, Massana R (2001) Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Appl Environ Microbiol. 67:2932-41

Dojka MA, Hugenholtz P, Haack SK, Pace NR (1998) Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. Appl Environ Microbiol. 10:3869-77

Dojka MA, Harris JK, Pace NR (2000) Expanding the known diversity and environmental distribution of an uncultured phylogenetic division of bacteria. Appl Environ Microbiol. 66:1617-21

Doolittle WF (1999a) Lateral genomics. Trends Cell Biol. 12:5-8

Doolittle WF (1999b) Phylogenetic classification and the universal tree. Science 5423:2124-9

Doerfler W (1996) A new concept in (adenoviral) oncogenesis: integration of foreign DNA and its consequences. Biochim Biophys Acta 1288:79-99

Dracheva S, Koonin EV, Crute JJ (1995) Identification of the primase active site of the herpes simplex virus type 1 helicase-primase. J Biol Chem. 23:14148-53

Edgell DR, Doolittle WF (1997) Archaea and the origin(s) of DNA replication proteins. Cell. 89:995-8

Edgell DR, Malik SB, Doolittle WF (1998) Evidence of independent gene duplications during the evolution of archaeal and eukaryotic family B DNA polymerases. Mol Biol Evol. 15:1207-17

Erauso G, Marsin S, Benbouzid-Rollet N, Baucher MF, Barbeyron T, Zivanovic Y, Prieur D, Forterre P (1996) Sequence of plasmid pGT5 from the archaeon Pyrococcus abyssi: evidence for rolling-circle replication in a hyperthermophile. J Bacteriol. 178:3232-7

Erzberger JP, Pirruccello MM, Berger JM (2002) The structure of bacterial DnaA: implications for general mechanisms underlying DNA replication initiation. EMBO J. 18:4763-73

Feder M, Pas J, Wyrwicz LS, Bujnicki JM (2003) Molecular phylogenetics of the RrmJ/fibrillarin superfamily of ribose 2'-O-methyltransferases. Gene. 302:129-38

Felsenstein J (1978) Cases in wich parsimony or compatibilty methods will be positively misleadind. Syst Zool. 27:401-410

Filée J, Forterre P, Sen-Lin T, Laurent J (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J Mol Evol. 54:763-773

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

Forterre P, Benachenhou-Lahfa N, Confalonieri F, Duguet M, Elie C, Labedan B (1992) The nature of the last universal ancestor and the root of the tree of life, still open questions. Biosystems. 28:15-32

Forterre P, Benachenhou-Lafha N, Labedan B (1993) Universal tree of life. Nature 362:795

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

Forterre P, Philippe H (1999) Where is the root of the universal tree of life? Bioessays 10:871-9

Forterre P, Bouthier De La Tour C, Philippe H, Duguet M (2000) Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. Trends Genet.16:152-4

Forterre P. (2002) The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol. 5:525-32

Forterre P (2002) A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. Trends Genet. 18:236-7

Forterre P, Brochier C, Philippe H (2002) Evolution of the Archaea. Theor Popul Biol.61:409-22

Galperin M, Koonin E (1999) Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. Genetica 106:159-70

Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol. 44:632-6

Galtier N, Tourasse N, Gouy M (1999) A nonhyperthermophilic common ancestor to extant life forms. Science 5399:220-1

Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol. 18:866-73

Garcia-Vallve S, Romeu A, Palau J (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 11:1719-25

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T (1989) Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA 86:6661-5

Golding GB, Gupta RS (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol Biol Evol. 12:1-6

Gorbalenya AE, Koonin EV, Wolf YI (1990) A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses. FEBS Lett. 1990 262:145-8

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8:49-62

Gray M (1998) Rickettsia, typhus and the mitochondrial connection. Nature 396:109-110

Gray M, Lang B (1998) Transcription in chloroplasts and mitochondria: a tale of two polymerases. Trends Microbiol. 6:1-3

Gray MW, Burger WG, Lang BF (1999) Mitochondrial evolution. Science 283:1476-1481

Guillot S, Caro V, Cuervo N, Korotkova E, Combiescu M, Persu A, Aubert-Combiescu A, Delpeyroux F, Crainic R (2000) Natural genetic exchanges between vaccine and wild poliovirus strains in humans. J Virol. 74:8434-43

Haeckel E, Generelle Morphologie der Organismen, Berlin, 1866.

Hannaert V, Saavedra E, Duffieux F, Szikora JP, Rigden DJ, Michels PA, Opperdoes FR (2003) Plant-like traits associated with metabolism of Trypanosoma parasites. Proc Natl Acad Sci USA 100:1067-71

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H (2001) Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res.28:11-22

Hedtke B, Borner T, Weihe A (1997) Mitochondrial and chloroplast phage-type RNA polymerases in Arabidopsis. Science 277:809-811

Hendrix R (1999) Evolution: the long evolutionary reach of viruses. Curr Biol. 9:R914-917

Hennig W (1966) Phylogenetic Systematics. Urbana Univ., Illinois, USA.

Highton PJ, Chang Y, Myers RJ (1990) Evidence for the exchange of segments between genomes during the evolution of lambdoid bacteriophages. Mol Microbiol. 8:1329-40

Hilario E, Gogarten JP (1995) Horizontal transfer of ATPase genes--the tree of life becomes a net of life. Biosystems. 31:111-9

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature 6884:63-7

Hughes A (2002) Origin and evolution of viral interleukin-10 and other DNA virus genes with vertebrate homologues. J Mol Evol 54:90-101

Hughes AL, Friedman R. Genome-wide survey for genes horizontally transferred from cellular organisms to baculoviruses. Mol Biol Evol. 20:979-87

Hodel AE, Gershon PD, Shi X, Quiocho FA (1996) The 1.85 A structure of vaccinia protein VP39: a bifunctional enzyme that participates in the modification of both mRNA ends. Cell. 85:247-56

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355-9

Iyer LM, Aravind L, Koonin EV (2001) Common origin of four diverse families of large eukaryotic DNA viruses. J Virol. 75:11720-34

Jackson MP, Newland JW, Holmes RK, O'Brien AD (1987) Nucleotide sequence analysis of the structural genes for Shiga-like toxin I encoded by bacteriophage 933J from Escherichia coli. Microb Pathog. 2:147-53

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801-6.

Jordan A, Reichard P (1998) Ribonucleotide reductases. Annu Rev Biochem.:71-98

Kakinuma Y, Igarashi K, Konishi K, Yamato I (1991) Primary structure of the alpha-subunit of vacuolar-type Na(+)-ATPase in Enterococcus hirae. Amplification of a 1000-bp fragment by polymerase chain reaction. FEBS Lett. 292:64-8

Kaneko T, Nakamura Y, Sato S, Asamizu E, Kato T, Sasamoto S, Watanabe A, Idesawa K, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A, Mochizuki Y, Nakayama S, Nakazaki N, Shimpo S, Sugimoto M, Takeuchi C, Yamada M, Tabata S (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. DNA Res. 31:331-8

Kelman Z, O'Donnell M (1995) Structural and functional similarities of prokaryotic and eukaryotic DNA polymerase sliding clamps. Nucleic Acids Res. 23:3613-20

Knopf CW (1998) Evolution of viral DNA-dependent DNA polymerases. Virus Genes. 16:47-58

Kobayashi I (1998) Selfishness and death: raison d'etre of restriction, recombination and mitochondria. Trends Genet. 14:368-74.

Koonin EV, Ilyina TV (1992) Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. J Gen Virol. 73:2763-6

Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol. 55:709-42

Kornberg A, Baker T (1992) DNA replication. Freeman and Company, NY

Krogh BO, Shuman S (2002) A poxvirus-like type IB topoisomerase family in bacteria. Proc Natl Acad Sci USA 99:1853-8

La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D (2003) A giant virus in amoebae. Science 299:2033

Lan R, Reeves PR (1996) Gene transfer is a major factor in bacterial evolution. Mol Biol Evol. 1996 13:47-55

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, Lemieux C, Sankoff D, Turmel M, Gray MW (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. Nature 387:493-7

Lawrence JG (1997) Selfish operons and speciation by gene transfer. Trends Microbiol. 9:355-9

Lawrence JG, Hatfull GF, Hendrix RW (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol. 184:4891-905

Leipe D, Aravind L, Grishin N, Koonin E (2000) The bacterial replicative helicase DnaB evolved from a RecA duplication. Genome Res. 10:5-16

Lerat E, Capy P (2001) Retrotransposons and retroviruses: analysis of the envelope gene. Mol Biol Evol. 161198-207.

Lewis SM, Wu GE (1997) The origins of V(D)J recombination. Cell. 88:159-62

Lipps G, Rother S, Hart C, Krauss G. A novel type of replicative enzyme harbouring ATPase, primase and DNA polymerase activity. EMBO J. 10:2516-25.

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C, Moreira D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature 6820:603-7

Margulis L (1971) Symbiosis and the evolution Sci Am. 225:48-57

Martin W, Muller M. (1998) The hydrogen hypothesis for the first eukaryote. Nature 6671:37-41

Martin IV, MacNeill SA (2002) ATP-dependent DNA ligases. Genome Biol. 3:R

Matte-Taillez O, Brochier C, Forterre P, Phillipe H (2002) Acrchaeal phylogeny based on ribosomal proteins. Mol Biol Evol 19:631-639

McClure MA (1991 Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol Biol Evol. 8:835-56

McDonald JF, Matyunina LV, Wilson S, Jordan IK, Bowen NJ, Miller WJ (1998) LTR retrotransposons and the evolution of eukaryotic enhancers. Genetica. 100:3-13

Medigue C, Rouxel T, Vigier P, Henaut A, Danchin A (1980) Evidence for horizontal gene transfer in Escherichia coli speciation. J Mol Biol 222:851-6

Mesyanzhinov VV, Robben J, Grymonprez B, Kostyuchenko VA, Bourkaltseva MV, Sykilinda NN, Krylov VN, Volckaert G (2002) The genome of bacteriophage phiKZ of Pseudomonas aeruginosa. J Mol Biol.317:1-19

Montague MG, Hutchison CA 3rd (2000) Gene content phylogeny of herpesviruses. Proc Natl Acad Sci USA. 97:5334-9

Moon-van der Staay SY, De Wachter R, Vaulot D (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature 6820:607-10

Moreira D (2000) Horizontal transfer of informational genes. Molecular Microbiology 35:1-5

Moreira D, Lopez-Garcia P (1998) Symbiosis between methanogenic archaea and delta-proteobacteri as the origin of eukaryotes: the syntrophic hypothesis. J Mol Evol. 47:517-30

Moreira D, Le Guyader H, Phillipe H (2000) The origin of red algae : implication for the evolution of photosynthetic plastids. Nature 405:69-72

Moreira D, Phillipe H (2001) Sure facts and open questions about the origin and evolution of photosynthetic plastids. Res Microbiol 152(9):771-80

Moreira D, Lopez-Garcia P (2002) The molecular ecology of microbial eukaryotes unveils a hidden world. Trends Microbiol. 10:31-8

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93:10268-73

Myllykallio H, Lopez P, Lopez-Garcia P, Heilig R, Saurin W, Zivanovic Y, Philippe H, Forterre P. (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. Science 5474:2212-5

Myllykallio H, Lipowski G, Leduc D, Filee J, Forterre P, Liebl U (2002) An alternative flavin-dependent mechanism for thymidylate synthesis. Science 297:105-107

Nagy M, Nagy E, Tuboly T (2002) Sequence analysis of porcine adenovirus serotype 5 fibre gene: evidence for recombination. Virus Genes 24:181-5

Naito T, Kusano K, Kobayashi I (1995) Selfish behavior of restriction-modification systems. Science 5199:897-9

Nesbo CL, Boucher Y, Doolittle WF Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J Mol Evol.53:340-50

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 6734:323-9

Nilsson AS, Haggard-Ljungquist E (2001) Detection of homologous recombination among bacteriophage P2 relatives. Mol Phylogenet Evol. 21:259-69

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 6784:299-304

Olsen GJ, Woese CR (1997) Archaeal genomics: an overview. Cell. 89:991-4

Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734-740

Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR Jr, Hendrix RW, Hatfull GF (2003) Origins of highly mosaic mycobacteriophage genomes. Cell. 113:171-82.

Peng X, Blum H, She Q, Mallok S, Brugger K, Garrett R, Zillig W, Prangishvili D (2001) Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. Virology 291:226-234

Philippe H (1993) MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res. 21:5264-72

Philippe H, Laurent J (1998) How good are deep phylogenetic trees? Curr Opin Genet Dev. 6:616-23

Philippe H, Adoutte A (1998) The molecular phylogeny of Eukaryota. In Evolutionnary relationships among protozoa. Pp 25-56 Kluwer, Dordrecht.

Philippe H, Forterre P (1999) The rooting of the universal tree of life is not reliable. J Mol Evol. 49:509-23

Poole A, Penny D, Sjoberg B (2000) Methyl-RNA: an evolutionary bridge between RNA and DNA? Chem Biol. 12:207-16

Poole A, Penny D, Sjoberg BM (2001) Confounded cytosine! Tinkering and the evolution of DNA. Nat Rev Mol Cell Biol. 2:147-51

Prangishvili D, Stedman K, Zillig W (2001) Viruses of the extremely thermophilic archaeon Sulfolobus. Trends Microbiol. 9:39-43

Raleigh EA, Wilson G (1986) Escherichia coli K-12 restricts DNA containing 5-methylcytosine. Proc Natl Acad Sci USA. 23:9070-4

Ravin V, Ravin N, Casjens S, Ford ME, Hatfull GF, Hendrix RW (2000) Genomic sequence and analysis of the atypical temperate bacteriophage N15. J Mol Biol. 299:53-73

Reanney DC (1974) On the origin of prokaryotes. J Theor Biol 48:243-51.

Repoila F, Tetart F, Bouet JY, Krisch HM (1994) Genomic polymorphism in the T-even bacteriophages. EMBO J. 13:4181-92

Roberts RJ, Macelis D (1997) REBASE-restriction enzymes and methylases. Nucleic Acids Res. 25:248-62

Robertson DL, Sharp PM, McCutchan FE, Hahn BH (1995) Recombination in HIV-1. Nature 374:124-6

Rocha EP, Danchin A. (2002) Base composition bias might result from competition for metabolic resources. Trends Genet. 18:291-4

Rohwer F, Edwards R (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. J Bacteriol. 184:4529-35

Sanger F, Ai GM, Barrel BG, Brown AR, Coulson JC, Fiddes CA, Hutchison PM, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 165:687-695

Schwartz RM, Dayhoff MO (1978) Origins of prokayotes, eukaryotes, mitochondria and chloroplasts. Science 199:395-403

Simon M, Zieg J, Silverman M, Mandel G, Doolittle R (1980) Phase variation: evolution of a controlling element. Science 209:1370-4.

Sintchak M, Arjara G, Kellogg B, Stubbe J, Drennan C (2002) The crystal structure of class II ribonucleotide reductase reveals how an allosterically regulated monomer mimics a dimer. Nat Struct Biol. 9:293-300

Slesarev AI, Stetter KO, Lake JA, Gellert M, Krah R, Kozyavkin SA (1993) DNA topoisomerase V is a relative of eukaryotic topoisomerase I from a hyperthermophilic prokaryote. Nature 364:735-7.

Song HK, Sohn SH, Suh SW (1999) Crystal structure of deoxycytidylate hydroxymethylase from bacteriophage T4, a component of the deoxyribonucleoside triphosphate-synthesizing complex. EMBO J. 18:1104-13

Sowers KR (1995) Restriction-modification system of methanogenic Archaea. Cold spring harbor laboratory press, Cold spring Harbor, NY

Spanopoulou E, Zaitseva F, Wang FH, Santagata S, Baltimore D, Panayotou G (1996) The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination. Cell 87:263-76

Spelbrink J, Li F, Tiranti V, Nikali K, Yuan Q, Tariq M, Wanrooij S, N G, Comi G, Morandi L, Santoro L, Toscano A, Fabrizi G, Somer H, Croxen R, Beeson D, Poulton J, Suomalainen A, Jacobs H, Zeviani M, Larsson C (2001) Human mitochondrial DNA deletions associated with mutations in the gene encoding Twinkle, a phage T7 gene 4-like protein localized in mitochondria. Nat Genet. 28:223-231

Strimmer K, von Haeseler (1996) Quartet puzzling : a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964-969.

Subramanian G, Koonin EV, Aravind L (2000) Comparative genome analysis of the pathogenic spirochetes Borrelia burgdorferi and Treponema pallidum. Infect Immun.68:1633-48.

Sumi M, Sato MH, Denda K, Date T, Yoshida M (1992) A DNA fragment homologous to F1-ATPase beta subunit was amplified from genomic DNA of Methanosarcina barkeri. Indication of an archaebacterial F-type ATPase. FEBS Lett. 314:207-10

Takemura M (2001) Poxviruses and the origin of the eukaryotic nucleus. J Mol Evol. 52:419-25

Teodoro JG, Branton PE (1997 Regulation of apoptosis by viral gene products. J Virol. 1997 71:1739-46

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-80

Torvisk T, Dundas ID (1978) Halophilic phage specific for Halobacterium salinarium. Pp609-615.In Energetics and structure of halophilic microorganism. Elsevier, Amsterdam.

Tsutsumi S, Denda K, Yokoyama K, Oshima T, Date T, Yoshida M (1991) Molecular cloning of genes encoding major two subunits of a eubacterial V-type ATPase from Thermus thermophilus. Biochim Biophys Acta. 1098:13-20

Van Etten JL, Meints RH. Giant viruses infecting algae. Annu Rev Microbiol. 53:447-94

Weeks CR, Ferretti JJ (1984) The gene for type A streptococcal exotoxin (erythrogenic toxin) is located in bacteriophage T12. Infect Immun. 46:531-6

Villarreal LP (1999) DNA virus contribution to host evolution. Academic Press, San Diego

Villarreal LP, DeFilippis VR (2000) A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. J Virol 74:7079-84

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 31:258-61

Waldor MK, Mekalanos JJ (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. Science. 272:1910-4

Williams BA, Hirt RP, Lucocq JM, Embley TM (2002)A mitochondrial remnant in the microsporidian Trachipleistophora hominis. Nature 418:865-9

Woese CR (1982) Archaebacteria. Scientific American. 94-106

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221-271.

Woese CR, Fox GE, Zablen L, Uchida T, Bonen L, Pechman K, Lewis BJ, Stahl D (1975) Conservation of Primary Structure in 16S rRNA. Nature 254: 83-85

Woese CR, Fox GE (1977) Phylogenetic structure of prokaryotic domain : the primary kingdoms. Proc Natl Acad Sci USA. 74:5088-5090

Woese CR, Olsen GJ (1986) Arcaebacterial phylogeny: perspectives on the urkingdoms. Syst Appl microbiol. 7:161-177

Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev. 64:69-114

Wu X, Guarino LA (2002) Autographa californica nucleopolyhedrovirus orf69 encodes an RNA cap (nucleoside-2'-O)-methyltransferase. J Virol.77:3430-40.

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 10:3353-62

Yang MY, Bowmaker M, Reyes A, Vergani L, Angeli P, Gringeri E, Jacobs HT, Holt IJ (2002) Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. Cell. 2002 111:495-505

Zgur-Bertok D (1999) Mechanisms of horizontal gene transfer. Folia Biol (Praha) 45:91-6.

Zillig W (1987) Eukaryotic traits in Archaebacteria. Could the eukaryotic cytoplasm have arisen from archaebacterial origin ? Ann N Y Acad Sci. 503:78-82.

Zillig W PD, Schleper C, Elferink M, Holz I, Albers S,, Janekovic D GD (1996) Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea. FEMS Microbiol Rev. 18:225-236

Zuckerkandl E, Pauling L (1965) Evolutinary divergence and convergence in proteins. 97-166 in V. Bryson et H.J. Vogel, eds Evolving Genes and Proteins. Academic Press, New York.

RESUME

Un faisceau de preuves tend à démontrer que les virus sont des éléments génétiques très anciens, d'une origine probablement antérieure à la divergence des trois domaines du vivant. Cette longue histoire suggère fortement que les virus aient pu jouer un rôle important dans l'évolution de leurs hôtes. Cette hypothèse est particulièrement pertinente en ce qui concerne les gènes viraux ayant des homologues cellulaires et soulève la question de leur relation phylogénétique.

Les génomes viraux codent pour diverses enzymes impliquées dans le métabolisme et la réplication de l'ADN. Les gènes correspondants sont très souvent phylogénétiquement éloignés de ceux de leurs hôtes ; par contre, quand ils sont étroitement apparentés, souvent, l'explication la plus probable indique que le gène cellulaire est d'origine virale. La situation est particulièrement intéressante dans le cas des mitochondries, où au moins trois enzymes cellulaires auraient été remplacées par des contre parties virales.

Ces propositions s'inscrivent dans la problématique plus générale de l'évolution et de l'origine des enzymes informationnelles. Nous montrons que les répartitions phylogénétiques de ces enzymes ne supportent pas l'hypothèse d'une double invention de l'ADN : une dans la lignée des Archéobactéries/Eucaryotes et une dans la lignée des Bactéries. Pour rendre compte de ces répartitions, il est plus vraisemblable d'imaginer de nombreux évènements de transferts horizontaux de gènes entre les trois domaines cellulaires du vivant, et entre cellule et virus, suivis, ou non, du remplacement non-homologue du gène initialement présent.

Ces travaux posent aussi clairement l'importance de l'échantillonnage de séquences utilisé : une meilleure connaissance de la biodiversité des virus et des êtres cellulaires pourra sans aucun doute éclaircir les points encore en suspens à l'issue de ce travail.