Que savons-nous de l'histoire évolutive des eucaryotes ? 1. L'arbre universel du vivant et les difficultés de la reconstruction phylogénétique

Hervé Philippe, Agnès Germot, Hervé Le Guyader, André Adoutte

Les eucaryotes forment un ensemble de lignées au sein duquel – avec les animaux, les plantes et les champignons – se retrouvent tous les grands groupes biologiques qui, pour la majorité d'entre nous, paraissent constituer l'essentiel de la diversité du vivant. Jusqu'à une période relativement récente, l'intérêt s'est presque exclusivement porté sur ces organismes multicellulaires de grande taille, donc facilement observables, source à la fois de notre alimentation et de notre plaisir esthétique, et, de surcroît, contenant notre propre espèce.

Trois notions se sont naturellement installées et ont, implicitement et explicitement, structuré les idées sur l'évolution des eucaryotes. Tout d'abord, on a considéré comme allant de soi que les eucaryotes « complexes » dérivaient des procaryotes, structurellement « simples » ; on a également pensé que l'essentiel de la durée de l'histoire du vivant était dévolu à la diversification des eucaryotes ; enfin, en guise de corollaire, on a admis que les distances évolutives étaient énormes entre les grands groupes d'eucaryotes.

Ces idées sont exprimées dans la majorité des arbres évolutifs, depuis celui, fameux, d'Haeckel de 1874 [1], jusqu'à celui du récent manuel, désormais classique, d'Alberts *et al.*, du moins dans ses premières éditions (1983, 1989) [2]. L'essentiel de l'arbre est occupé par des branches de grande taille, profondément séparées les unes des autres, et correspondant aux plantes, aux champignons, et aux animaux. Sur chacune des branches, des rameaux émergent régulièrement, reflétant la notion d'une « complexité croissante ». Pour Haeckel, chez qui la distinction entre procaryotes et eucaryotes n'est pas encore faite, on trouve à la base de l'arbre une mince couche de « monères » dans laquelle plantes, animaux et protistes plongent directement leurs racines. Pour Alberts et al., un niveau inférieur, à nouveau mince, de bactéries est suivi de protistes qui donnent eux-mêmes naissance aux différents groupes de multicellulaires.

En fait, si la découverte des unicellulaires date du XVII^{e} siècle et si une description floue de l'ensemble des « microbes » date du XIX^{e} siècle [3], la distinction entre procaryotes et eucaryotes est relativement récente en biologie et est étroitement liée aux progrès de la microscopie, photonique d'abord, puis surtout électronique. Chatton [4], puis Stanier et Van Niel [5] vont définir les principales différences entre les deux grands types d'organisation cellulaire. Chez les procaryotes, on constate l'absence de réseaux membranaires internes – et en particulier autour de l'ADN, c'est-à-dire l'absence d'un noyau bien délimité –, ainsi que l'absence d'organites cytoplasmiques individualisés ; la division du matériel génétique s'y fait non par mitose, mais par un processus de ségrégation dans lequel l'ADN bactérien demeure lié à l'enveloppe cellulaire. Enfin, chez les procaryotes, une paroi mucopeptidique peut fournir une armature externe à la cellule.

Dans le même temps, l'unité fondamentale de tous les types cellulaires est mise en évidence par la biochimie

et par la biologie moléculaire naissante : identité des grands mécanismes biochimiques, en particulier de certaines voies centrales du métabolisme telle la glycolyse, mais surtout identité des mécanismes de stockage et d'expression de l'information génétique. Ces similitudes, jointes à la découverte de l'universalité du code génétique, ont donc conduit à admettre une origine évolutive commune à tous les types cellulaires. Cette lointaine parenté admise, il restait à se représenter la succession des étapes menant à la diversification du vivant.

Mais pour pouvoir comparer entre eux des organismes et aller jusqu'à l'établissement de leurs liens de parenté (c'est-à-dire reconstruire leur phylogénie), on doit disposer de caractères homologues (c'est-à-dire hérités d'un ancêtre commun) qui se présentent sous des états différents chez ces diverses espèces. Cela n'est possible que chez des organismes qui ont le même plan d'organisation. Ainsi, chez les vertébrés, il n'est pas difficile de reconnaître l'homologie des grandes parties du squelette et de faire des hypothèses sur les modifications survenues dans chacune de ces parties, ce qui permet à la fois de définir les grands groupes (poissons osseux, amphibiens, oiseaux, mammifères...), et d'imaginer l'ordre dans lequel les modifications se sont réalisées, ce qui permet de retracer les ordres d'apparition successives des différents groupes. Cette approche a été rationalisée et systématisée par Hennig [6], et est connue sous le nom de cladistique. A défaut de l'existence de tels caractères, on ne peut qu'identifier des groupes bien différents les uns des autres, au sein desquels les individus sont clairement apparentés, mais entre lesquels on ne sait pas déduire les filiations. C'est ainsi que les grandes subdivisions taxonomiques ont été réalisées et qu'il est facile de distinguer un annélide d'un arthropode ou d'un mollusque par exemple. Mais quels sont les deux groupes qui sont les plus proches entre eux par rapport au troisième ? Le problème est encore plus aigu quand il s'agit de comparer des protistes entre eux, et que dire des bactéries ! Il n'est pas surprenant, dans ces conditions, que le principe largement utilisé dans les phylogénies « traditionnelles » ait été celui de la « complexité croissante ». En l'absence de caractères homologues clairement analysables, on a eu recours à l'hypothèse (souvent raisonnable) que les organismes d'apparence simple avaient émergé, au cours de l'évolution, avant les organismes d'apparence complexe. Quant aux comparaisons entre organismes encore plus distants, on avait encore moins de moyens de définir leurs parentés et, pendant longtemps, surtout dans les pays anglo-saxons, on a utilisé le système des cinq règnes de Whittaker [7] (animaux, plantes, champignons, protistes, bactéries) entre lesquels les filiations demeuraient très incertaines. C'est à ce niveau que s'est situé l'apport principal de la phylogénie moléculaire : en permettant l'accès au génome et non plus seulement à la morphologie, la biologie moléculaire a permis d'accéder aux gènes homologues qui existent même entre organismes très distants. Ces gènes fournissent un ensemble de « caractères » sur lesquels se fondent les phylogénies. Par cette approche, la vision somme toute satisfaisante de l'histoire de la vie, esquissée jusqu'au milieu du XX^{e} siècle, va être sérieusement renouvelée.

Dans cet article, avant d'exposer la nouvelle vision de l'arbre des eucaryotes, nous allons procéder à un bref rappel des principes des phylogénies moléculaires. Nous présenterons les diverses difficultés existant tant dans l'analyse des données que dans l'interprétation des résultats. Ce point est d'autant plus important que les problèmes sont exacerbés pour les études évolutives à grande distance, puisqu'il s'agit de reconstituer une histoire s'étant déroulée sur plusieurs milliards d'années. L'étude de l'arbre universel du vivant permettra d'une part d'illustrer ces difficultés et d'autre part d'analyser le problème de l'origine des eucaryotes. De plus, des résultats récents ont permis de mettre en évidence des similitudes troublantes existant entre des gènes typiquement eucaryotes, comme celui de la tubuline, et des gènes procaryotes. Ainsi, le fossé qui sépare les procaryotes des eucaryotes du point de vue morphologique ne semble pas aussi marqué du point de vue moléculaire. Dans un deuxième article, nous présenterons la phylogénie à l'intérieur des eucaryotes. Celle-ci nous permettra de proposer des scénarios évolutifs pour quelques caractéristiques des cellules eucaryotes, comme les mitochondries, les introns ou l'appareil de Golgi.

Phylogénie moléculaire

Rappel des principes

Les espèces actuelles contiennent dans leur génome des séquences héritées d'un ancêtre commun. De telles séquences sont dites homologues. Après spéciation, elles ont évolué indépendamment les unes des autres. Au cours du temps, elles ont subi des mutations qui ont modifié le génome d'un individu. Il peut s'agir soit de mutations défavorables à l'organisme (pour l'essentiel éliminées par la sélection naturelle), soit de mutation neutres ou favorables. Il faut que ces mutations soient fixées à l'intérieur de l'espèce, c'est-à-dire que tous les individus acquièrent ces mutations, pour qu'elles soient observables aujourd'hui. Ainsi, à différents endroits d'un gène, des changements, des insertions et des délétions de nucléotides s'introduisent au fil des générations.

Si un gène homologue n'a pas trop divergé, on peut reconnaître des similitudes entre les séquences de diverses espèces. Cette condition est un prérequis indispensable à toute analyse phylogénétique. En effet, deux séquences doivent être suffisamment similaires pour qu'on puisse les supposer homologues, c'est-à-dire être héritées d'un ancêtre commun. L'alignement (c'est-à-dire l'optimisation des similitudes par l'introduction de « gaps », correspondant à d'hypothétiques événements d'insertion ou de délétion de nucléotides) est ensuite nécessaire pour trouver des sites homologues, c'est-à-dire des nucléotides ou des acides

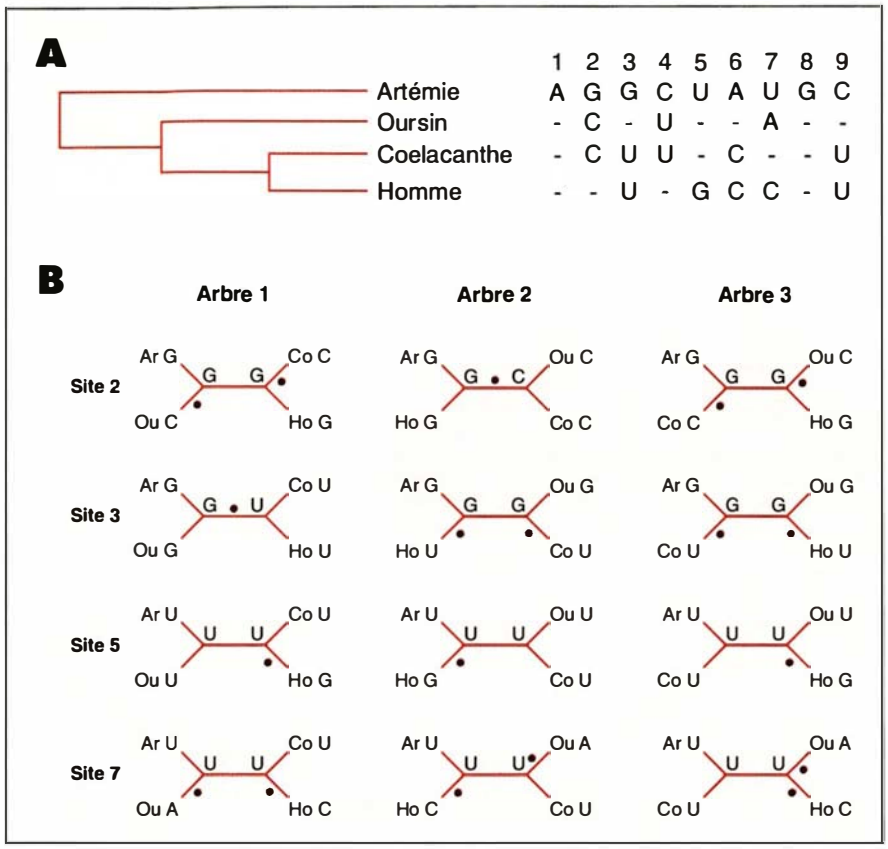

Figure 1. ***Les bases de la phylogénie moléculaire.*** *Soit 3 espèces à classer: un oursin (échinoderme), le coelacanthe (actinistien) et l'homme (mammifère). Comme il n'est pas possible de connaître le sens des substitutions nucléotidiques (représentées ici par un point bistre), les phylogénies moléculaires ne permettent pas de positionner l'ancêtre commun à ces 3 espèces, et il faut disposer d'une quatrième espèce comme groupe extérieur, ici l'artémie (arthropode), pour raciner correctement l'arbre.* ***A.*** *exemples de sites extraits des séquences alignées d'ARNr 18S. L'arbre représenté sur le côté est celui trouvé à l'aide des séquences complètes. Il est en accord avec les données de la zoologie, c'est-à-dire que le coelacanthe est le groupe-frère de l'homme.* ***B.*** *Étude de quelques sites et situation des étapes évolutives correspondantes sur les trois arbres non racinés possibles (Ar : artémie ; Ou : oursin ; Co : coelacanthe ; Ho : homme). L'arbre 1 est l'arbre A non raciné. Les sites 1 et 8, présentant le même état chez les 4 espèces, ne sont pas informatifs. Le site 2 soutient l'arbre 2: un seul événement évolutif explique les états des sites, tandis que deux événements sont nécessaires pour les 2 autres arbres (même explication pour le site 4). Le site 3 est contradictoire avec le site 2, car il soutient l'arbre 1 (même explication pour les sites 6 et 9). Le site 5 n'est pas informatif : le changement d'état est toujours sur une branche terminale. Au site 7, on a des états multiples. Ce site n'est pas informatif, car chaque arbre présente le même nombre de changement d'états. Si on fait, pour chacun des trois arbres, la somme des changements évolutifs correspondant aux 9 sites informatifs du tableau, on obtient : arbre 1 : 10 pas ; arbre 2 : 11 ; arbre 3 : 13. L'arbre 1 est donc le plus parcimonieux, et c'est celui qui sera retenu. Le retour sur les sites permet de donner une interprétation des événements évolutifs. Ainsi, les sites 3, 6 et 9 correspondent à des homologies, car les états dérivés sont hérités d'un ancêtre commun. Les sites 2 et 4 correspondent à des convergences évolutives, car les états dérivés ont été acquis indépendamment.*

aminés placés en correspondance les uns sous les autres. Les phylogénies sont alors construites grâce à l'analyse comparée de ces sites, de manière identique à ce qui est fait en anatomie comparée. Un site moléculaire est donc analogue à un caractère morphologique.

Les principes qui viennent d'être très rapidement résumés s'appliquent de manière identique à des intervalles évolutifs extrêmement divers, depuis la phylogénie à l'intérieur d'une espèce jusqu'à celle de l'ensemble du vivant. Les gènes dont la séquence est conservée sont utilisés pour les phylogénies à grande distance tandis que ceux dont la séquence est plus variable le sont pour celles à petite distance. Il existe en effet un grande diversité de vitesse d'évolution entre différentes portions du génome qui reflète l'intensité des contraintes fonctionnelles qui pèsent sur les gènes correspondants. Par exemple, les histones figurent parmi les protéines les plus conservées au plan évolutif en raison des interactions nombreuses qu'elles entretiennent entre elles et avec l'ADN. Il en est de même pour les tubulines et, d'une façon générale, pour les protéines ayant des interactions multiples avec leur substrat et avec d'autres protéines. A l'inverse, les apolipoprotéines, dont la principale contrainte est de demeurer hydrophobes, ont fixé 1000 fois plus de mutations dans le même intervalle évolutif que l'histone H3! A l'extrême, les pseudogènes et les séquences répétées non codantes fixent les mutations au taux maximal et conviennent pour la comparaison d'espèces très proches voire à l'intérieur d'une même espèce. C'est sur ce dernier type de séquences qu'on se fonde pour effectuer de la phylogénie intraspécifique, notamment pour celle qui se situe au plus petit intervalle évolutif concevable, l'établissement de paternité.

Le préalable à toute reconstruction phylogénétique est de disposer de séquences alignées présentant une variabilité raisonnable entre les espèces étudiées. Les deux méthodes les plus utilisées pour reconstruire l'arbre retraçant les relations de parenté entre les espèces actuelles, la

méthode de distances et la méthode de parcimonie, vont être succinctement présentées (pour une description détaillée des méthodes de reconstruction phylogénétique, voir [8 - 10]).

Pour le gène homologue choisi, on calcule les distances évolutives en dénombrant les différences existant entre deux séquences. En répétant cette opération pour toutes les paires d'espèces, on obtient une matrice de distances. Intuitivement, il est raisonnable de supposer que les deux espèces ayant la plus petite distance sont aussi les plus proches parentes. La méthode de l'UPGMA [11], qui se fonde sur ce principe, a été utilisée dès les années 60. Elle consiste à agglomérer les deux espèces les plus proches puis de proche en proche à agréger les espèces de plus en plus divergentes. La méthode ainsi décrite nécessite l'hypothèse de la constance du taux de substitutions dans les différentes branches de l'arbre afin de refléter les vraies relations de parenté. Néanmoins, comme cette hypothèse est biologiquement très discutable [12], de nombreuses méthodes de distances s'en affranchissant ont été mises au point. Elles reviennent à construire un arbre reflétant d'aussi près que possible les distances de la matrice de départ en allongeant certaines branches et en en raccourcissant d'autres. Les inégalités de vitesse de substitutions qui ont eu lieu dans les différentes lignées sont ainsi respectées au mieux.

Au lieu d'étudier globalement toutes les substitutions qui se sont produites entre deux espèces sur l'ensemble des sites, la méthode de parcimonie consiste à étudier toutes les substitutions ayant eu lieu à un site donné chez l'ensemble des espèces et cela pour l'ensemble des sites de la séquence. On recherche d'abord les singularités qui unissent des séquences par deux puis en groupes emboîtés de taille croissante, chaque groupe étant caractérisé par un spectre de substitutions qui lui est propre (des caractères « dérivés partagés » uniques à ce groupe) ; c'est le principe de l'approche cladistique, initialement développée pour l'analyse des caractères morphologiques [6] et qui est aussi applicable aux caractères moléculaires.

Si des événements mutationnels uniques se produisaient dans les différentes lignées évolutives étudiées, c'est-à-dire si toutes les mutations qui se sont fixées avaient affecté des nucléotides différents et ne les avaient affecté chacun qu'une fois, les deux types de méthode donneraient des résultats identiques. Il n'y aurait pas de difficulté à retracer les filiations évolutives. La réalité est bien différente ! On sait en effet que des substitutions se produisent de manière récurrente au même site. Ces événements vont brouiller le message phylogénétique ; la *figure 1* présente différents sites nucléotidiques et l'interprétation phylogénétique qui en est faite. Suivant les cas, le partage d'un même état de caractère à un site donné est interprété soit comme un état homologue – issu d'un même événement évolutif, ce qui constitue l'information phylogénétique –, soit comme des convergences – issues d'événements évolutifs différents, des substitutions multiples survenues au même site, ce qui constitue le « bruit » phylogénétique –. Ainsi, non seulement les substitutions multiples peuvent détruire le message phylogénétique (par exemple, une réversion), mais aussi elles peuvent générer de mauvais regroupements d'espèces (par exemple, une convergence). Elles constituent donc un des problèmes majeurs des phylogénies moléculaires. L'aptitude des méthodes de reconstruction à s'affranchir des événements multiples est l'objet de recherches et de discussions intenses dans la communauté des phylogénéticiens. Avant d'étudier l'arbre universel du vivant, nous allons présenter ce que nous estimons être les principaux écueils des phylogénies moléculaires.

Quelques problèmes

Le problème des substitutions multiples a été étudié d'un point de vue théorique dès les années 1970. Dès lors, on a cherché à connaître dans quelles conditions une méthode de reconstruction allait converger vers le bon résultat, « l'arbre vrai », à mesure que la quantité de séquences utilisées augmentait. Une méthode qui retrouve « l'arbre vrai » est dite consistante. Le résultat le plus spectaculaire a été trouvé par Felsenstein [13]. En effet, dans un cas simple à quatre espèces, il a montré que la méthode de parcimonie n'est pas consistante si deux des espèces évoluent beaucoup plus vite que les autres. En fait, quand deux espèces évoluent rapidement, la probabilité qu'apparaissent de faux caractères dérivés partagés à cause de convergences est élevée ; et il suffit que la fréquence de telles convergences soit supérieure à la fréquence des vrais caractères dérivés partagés pour que la méthode de parcimonie regroupe ces deux espèces dans l'arbre phylogénétique. Ce phénomène désormais célèbre est connu sous le nom « d'attraction des longues branches ». Pour s'affranchir d'un tel phénomène, les théoriciens se sont focalisés sur la mise au point de méthodes consistantes. Néanmoins, aucune de ces nouvelles méthodes n'a réussi à s'imposer, soit à cause d'un besoin de temps calcul démesuré (maximum de vraisemblance) soit à cause d'hypothèses discutables sur les processus de substitutions des nucléotides (parcimonie évolutive).

La *figure 2* présente une illustration pratique de l'attraction des longues branches. En effet, on observe qu'il suffit de changer les espèces représentant les groupes taxonomiques étudiés, ici des sous-ordres ou des ordres de mammifères, pour changer complètement la phylogénie inférée. Dans les deux arbres présentés, les résultats, bien que contradictoires, sont tous les deux statistiquement solides, confirmant que ce problème n'est pas dû à l'utilisation d'un trop petit nombre de nucléotides, mais à la non-consistance de la méthode utilisée. Le *bootstrap* [14], test statistique le plus fréquemment employé en phylogénie, a été utilisé dans ce cas. Son principe consiste à générer un corps de données de même taille que le corps de données initial en réalisant un tirage avec remise des sites. Cela revient à donner des poids aléatoires aux différents sites. Pour le

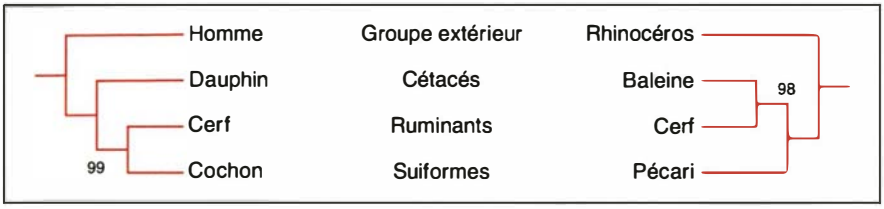

Figure 2. ***Attraction des longues branches et échantillonnage taxonomique.*** *Pour résoudre les relations phylogénétiques entre les cétacés et deux groupes d'artiodactyles [17], les ruminants et les suiformes, on a utilisé des séquences de cytochrome b et seulement quatre espèces, une seule par groupe monophylétique. En changeant la représentation taxonomique, il est possible de modifier complètement l'inférence phylogénétique, ceci par la méthode aussi bien de parcimonie que de distances. Ce phénomène est simplement dû au fait que les espèces sélectionnées évoluent à des vitesses différentes. Par exemple, le cerf et le cochon, évoluant moins vite que l'homme et le dauphin, se trouvent regroupés à cause de l'attraction des longues branches. Il est à noter que les phylogénies contradictoires sont soutenues par de fortes valeurs de bootstrap (98 et 99 %).*

corps de données ainsi constitué, on construit un arbre phylogénétique. Cette opération est répétée un grand nombre de fois et en conséquence, génère un grand nombre d'arbres. Pour un groupe donné d'espèces, la valeur de bootstrap, c'est-à-dire le nombre de fois où ce groupe est monophylétique parmi tous ces arbres, permet de mesurer la solidité de ce regroupement. Par exemple, sur la *figure 2*, le cerf et la baleine ont été regroupés dans 98 % des arbres obtenus après tirage aléatoire des sites.

En fait, ces deux espèces sont regroupées parce que leur gène a évolué plus vite. Il n'existe malheureusement que peu de solutions pour contourner cet obstacle. Une idée intéressante [15] consiste à ajouter des espèces dans le corps de données afin de « casser les grandes branches ». D'un point de vue empirique, il semble que cette approche soit assez efficace [16, 17], mais elle n'a pas, pour l'instant, été étayée théoriquement [18, 19]. En revanche, il a été mathématiquement démontré que l'utilisation de séquences qui ont un faible taux de substitutions permet d'augmenter la consistance des méthodes de reconstruction car elle réduit la probabilité d'utiliser des sites où se sont produites des substitutions multiples [18]. Cette proposition est toutefois très difficile à appliquer.

A priori, les gènes régissant les processus fondamentaux du fonctionnement cellulaire doivent avoir une vitesse d'évolution très lente. Ces mécanismes ubiquistes sont en particulier ceux de la réplication, de la transcription, de la traduction et des grandes voies métaboliques. Le facteur d'élongation EF-1α intervient lors de la traduction en fixant l'aminoacyl-ARNt aux ribosomes. La similitude moyenne entre les séquences d'EF-1α d'archébactéries et d'eucaryotes est de 50 %, suggérant que ce gène évolue très lentement. La réalité est bien plus compliquée que cela, ainsi que l'illustre la *figure 3*. En effet, on observe que le nombre de différences entre deux espèces (axe des Y) atteint un plateau aux alentours de 225 tandis que le nombre de substitutions qui se sont produites (axe des X) entre les deux mêmes espèces continue de croître jusqu'à 450. Ce phénomène, appelé saturation, est dû aux nombreuses substitutions qui se sont produites de manière récurrente au même site. En fait, un gène qui évolue lentement est un gène dont de nombreux sites sont invariants au cours du temps [20, 21], et non pas un gène dont les sites pris individuellement évoluent lentement. Ce phénomène est facilement analysable dans la mesure où il est connu que chez les protéines certains acides aminés ont un rôle crucial dans le maintien de la structure tridimensionnelle et dans le site actif. En revanche, pour les phylogénies, cela implique qu'il est quasiment impossible de trouver des séquences évoluant très lentement, comme le recommandent les théoriciens.

La *figure 3* illustre aussi le problème de l'inconsistance des reconstructions phylogénétiques. Le gène codant pour la thésaurine du xénope est issue d'une duplication du gène EF-1α. Cette protéine est très abondante dans les oocytes prévitellogéniques et permet le stockage des aminoacyl-ARNt [22]. Ainsi, sa fonction est légèrement différente de celle des EF-1α. Ce changement de fonction a contribué à une accélération de la vitesse d'évolution du gène de la thésaurine. La position phylogénétique de la thésaurine, à la base des eucaryotes, se trouve donc être est erronée, puisque ce gène est typiquement animal. En effet, une insertion de 12 acides aminés, présente uniquement chez les animaux et les champignons [23], se retrouve aussi chez la thésaurine, indiquant que la duplication lui ayant donné naissance s'est produite très tard dans l'histoire des eucaryotes. En fait, la longue branche de la thésaurine est « attirée » par la longue branche des archébactéries. On voit donc bien comment un des gènes les plus conservés évolutivement peut être saturé en substitutions et générer des arbres inconsistants.

En conclusion, la **qualité** des séquences utilisées, c'est-à-dire un faible taux de substitutions, est une condition nécessaire et suffisante pour reconstruire l'histoire évolutive dans le sens que l'on n'obtient pas de mauvais regroupements d'espèces. Néanmoins, il faut une certaine **quantité** de séquences pour résoudre toutes les relations de parenté. Par exemple, dans le cas des relations entre artiodactyles et cétacés *(figure2)*, les phylogénies obtenues avec beaucoup d'espèces ne sont pas fausses mais elles ne sont pas résolues [17]. Nous avons récemment modélisé l'évolution des valeurs de bootstrap en fonction du nombre de nucléotides [24], ce qui permet, moyennant l'hypothèse de l'existence d'une horloge moléculaire, d'esti-

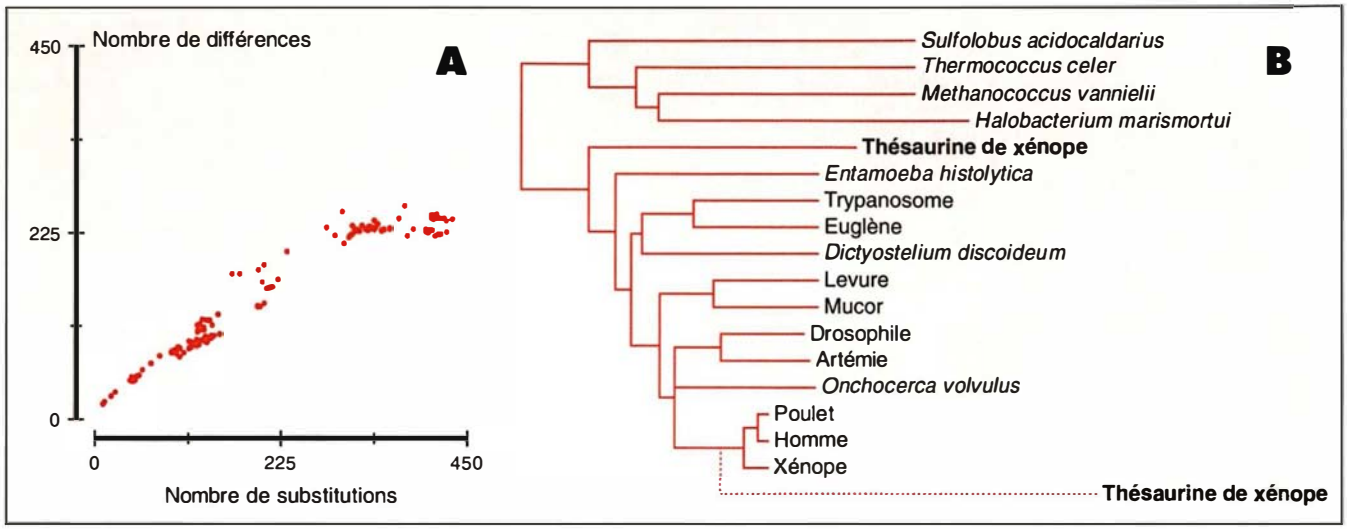

Figure 3. ***Substitutions multiples et saturation.*** ***A.*** *mise en évidence de la saturation. Afin de tester la qualité des informations à notre disposition, c'est-à-dire la fréquence des substitutions multiples, on peut construire un diagramme portant, pour chaque couple d'espèces, le nombre de substitutions calculées sur l'arbre obtenu par parcimonie, en fonction du nombre de différences observées sur les séquences [52]. On remarque que la courbe obtenue présente deux parties. Lorsque les distances ne sont pas très élevées, on observe une bonne corrélation entre les deux nombres ; les séquences sont vraiment informatives. Au-delà d'une certaine valeur du nombre des substitutions (environ 250), la courbe est horizontale, c'est-à-dire qu'il n'y a plus de corrélation entre les données de départ et les distances inférées. On dit que les données sont saturées, et ne sont plus pertinentes d'un point de vue phylogénétique. Ceci est dû à un taux élevé de substitutions qui brouillent le message phylogénétique.* ***B.*** *Les facteurs d'élongation forment une famille multigénique dont la phylogénie a été obtenue par une méthode de distances. Le groupe extérieur est formé par des archébactéries. On remarque notamment la monophylie des archébactéries, des Euglenozoa, des champignons et des métazoaires et l'émergence successive de différents de protistes. La thésaurine, protéine spécifique du xénope [22], sort à une place inattendue, étant donné qu'elle se trouve à la base des gènes eucaryotes. En fait, la branche conduisant à la thésaurine est extrêmement longue, en raison d'une grande vitesse d'évolution, sans doute corrélée à un changement de fonction de la molécule. Par le phénomène d'attraction des longues branches, celle-ci s'enracinera préférentiellement à la base de l'arbre. Il s'agit là d'un effet artéfactuel majeur sur la topologie, occasionné par des vitesses d'évolution différentes.*

mer le pouvoir résolutif des phylogénies moléculaires [25] *(figure 4)*. Quand le temps entre deux spéciations est très court, il faut un nombre considérable de nucléotides pour résoudre l'ordre d'émergence des taxons. A l'inverse, quand le temps entre deux spéciations est très long, les phylogénies moléculaires donnent un résultat solide, même avec des séquences courtes. Néanmoins, même s'il n'est pas possible de résoudre un problème phylogénétique, l'approche moléculaire fournit un intervalle de temps maximum durant lequel les spéciations se sont produites, ce qui apporte une information importante sur les processus évolutifs, comme, par exemple, sur les phénomènes de radiation adaptative [26].

Un dernier point mérite d'être soulevé. Pour l'instant, on a toujours interprété l'arbre phylogénétique obtenu à partir d'un gène homologue comme étant celui des espèces possédant les séquences de ce gène. Or, il est tout à fait possible qu'un tel arbre n'ait qu'un lointain rapport avec la phylogénie des espèces *(figure 5)*. En effet, si le gène utilisé appartient à une famille multigénique, il est difficile de différencier les spéciations des duplications. Pour qu'une phylogénie de gènes correspondent à une phylogénie d'espèces, il faut impérativement que les gènes homologues considérés soient orthologues. En d'autres termes, il faut que ces gènes n'aient acquis leur autonomie évolutive qu'après un événement de spéciation. Dans ce cas, ils appartiennent obligatoirement à des génomes d'organismes différents. Par opposition, des gènes sont dits paralogues lorsqu'ils ont acquis leur autonomie évolutive après une duplication génique. Des gènes homologues d'un même génome sont donc obligatoirement paralogues. Cet aspect peut poser des problèmes pour les phylogénies d'espèces (voir ci-dessous), mais l'étude des gènes paralogues est un élément de base pour comprendre l'évolution des génomes.

Malgré les diverses limitations que nous venons de présenter succinctement, la phylogénie moléculaire a démontré son utilité à des niveaux très différents de la hiérarchie taxonomique mais c'est sans doute aux échelles très élevées, celle de l'ensemble du vivant ou de grands

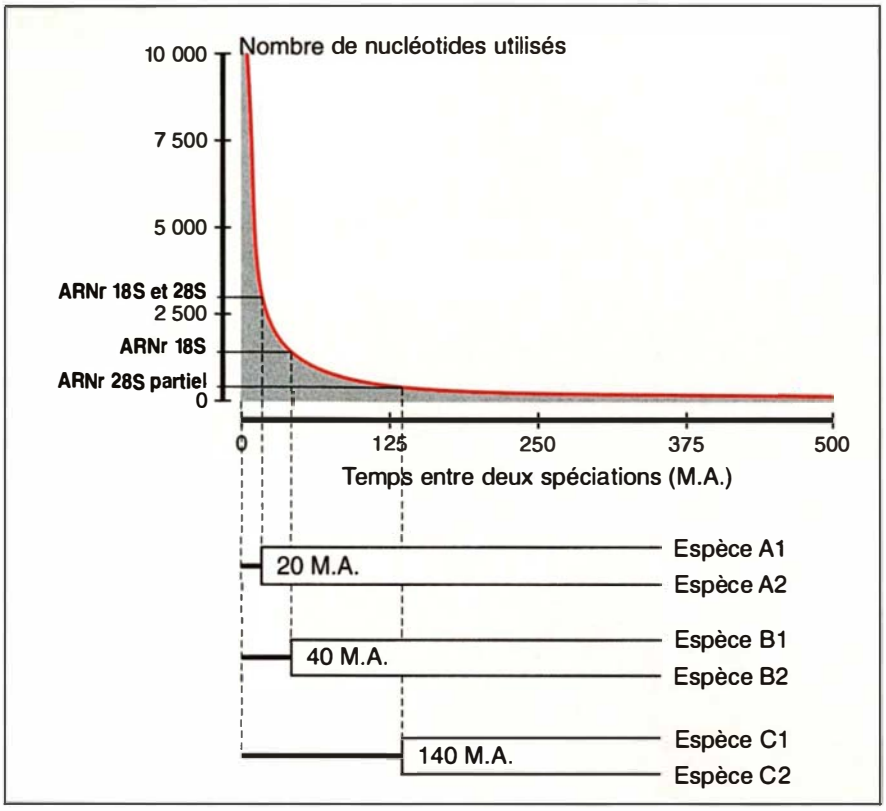

Figure 4. ***Pouvoir résolutif des phylogénies moléculaires.*** *On a représenté la relation entre le nombre de nucléotides nécessaire pour résoudre les ordres de branchement de manière statistiquement significative et le temps qui s'est écoulé entre ces branchements. Le diagramme a été obtenu dans le cas de la phylogénie des principaux phylums animaux [25]. La zone en gris est celle où la quantité d'information disponible ne permet pas de résoudre la phylogénie. Nous avons estimé que des séquences partielles d'ARNr 28S (environ 400 nucléotides) ne peuvent résoudre que les divergences séparées par plus de 140 millions d'années, l'ARNr 18S complet que celles séparées par plus de 40 millions d'années et les ARNr 18S et 28S que celles séparées par plus de 20 millions d'années. C'est ce que nous avons schématisé en forme d'arbre sous le graphe. Pour arriver à une résolution d'un million d'années, il faudrait avoir des données équivalentes à 40 fois celles de l'ARNr 18S pour toutes les espèces étudiées ce qui implique un effort de séquençage inimaginable.*

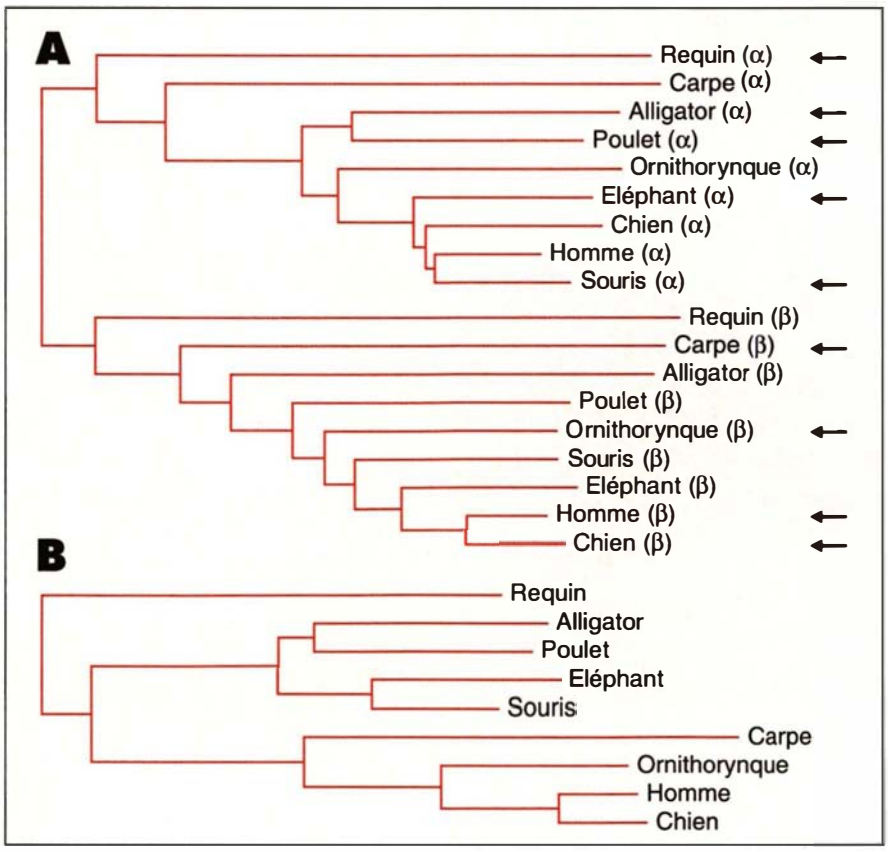

Figure 5. ***Gènes orthologues et gènes paralogues. A.*** *Phylogénie des familles des hémoglobines α et β. Le requin ayant déjà une hémoglobine tétramérique, la duplication qui a donné naissance aux deux gènes est plus ancienne que la divergence des poissons à mâchoires (gnathostomes).On retrouve à l'intérieur de chacune des familles la phylogénie d'espèces, étant donné que les gènes sont alors orthologues. Les quelques différences sont probablement dues au fait que les séquences utilisées sont courtes (environ 150 acides aminés) ;* ***B.*** *Phylogénie obtenue si des erreurs sont faites sur la détermination des paralogies et des orthologies. On a pris l'hémoglobine α pour le requin, l'alligator, le poulet, l'éléphant et la souris. On a choisi l'hémoglobine β pour la carpe, l'ornithorynque, l'homme et le chien. La phylogénie est aberrante si l'on suppose qu'elle représente celle des espèces, car les nœuds correspondent parfois à des duplications de gènes, parfois à des spéciations.*

ensembles de procaryotes et d'eucaryotes qu'elle a apporté les résultats les plus novateurs en raison, comme on l'a dit plus haut, de la difficulté ou de l'impossibilité des autres approches à disposer de caractères analysables. Pour reconstruire une phylogénie universelle, il faut donc disposer de gènes orthologues ubiquistes qui ont évolué lentement. Les gènes codant pour les grands ARN ribosomiques remplissent ces conditions et ont été très tôt utilisés pour établir les relations de parenté entre espèces. L'effort expérimental a surtout porté sur l'ARNr de la petite sous-unité du ribosome dont le coefficient de sédimentation varie de 16S chez les procaryotes à 18S chez les eucaryotes et dont la taille est comprise généralement entre 1 500 et 2 000 nucléotides. Pour l'instant, l'ARNr de la grande sous-unité (23S à 28S) a été moins séquencé, mais les résultats obtenus sont remarquablement convergents. Nous allons donc présenter les phylogénies obtenues avec l'ARNr de la petite sous-unité (que nous appellerons 18S), qui ont été initiées avec brio par Woese et ses collaborateurs [27] dans les années 70.

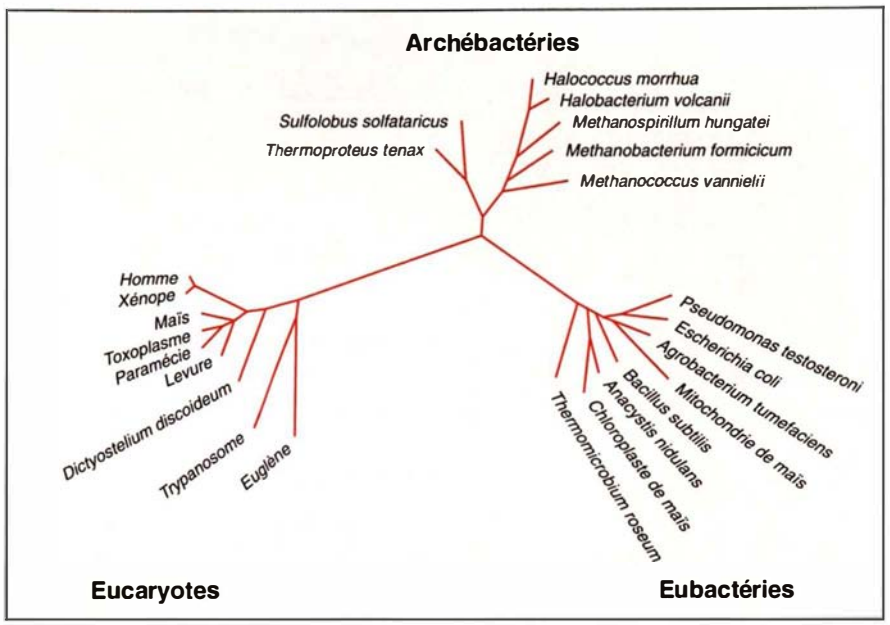

Figure 6. ***Arbre universel du vivant.*** *Les trois « Empires » constituant le monde vivant sont bien visibles. On remarque que les distances évolutives entre eucaryotes pluricellulaires sont très faibles comparativement à celles existant entre certains organismes unicellulaires (d'après [28] avec modifications). Cet arbre a été construit à partir de séquences d'ARN ribosomique 18S en utilisant une méthode de distances.*

Arbre universel du vivant

L'apport des ARN ribosomiques

L'analyse comparée de « catalogues d'oligonucléotides » issus de la digestion de l'ARNr 18S, première méthode utilisée avant la mise au point des techniques de séquençage rapide, apporte une conclusion surprenante [27] : le vivant ne se subdivise pas en deux super-ensembles, les procaryotes et les eucaryotes mais en trois. Les procaryotes se subdivisant eux-mêmes en deux ensembles de même rang, les archébactéries et les eubactéries (désormais appelées Archaea et Bacteria et que nous abrégerons par A et B sur les schémas, en réservant E pour les Eukarya ou eucaryotes). Un arbre de distance datant de 1987 [28], fondé sur des séquences complètes d'ARNr 18S, est présenté sur la *figure 6*. Les principales inférences que l'on peut en tirer ont déjà été abondamment discutées [28-30]. Rappelons-les succinctement en trois points.

a. Un ensemble de procaryotes se regroupent en un rameau distinct de celui qui regroupe les bactéries plus « communes » (*E. coli*, *B. subtilis*, cyanobactéries, etc.). Ces bactéries ont d'abord été découvertes dans des niches écologiques « extrêmes » (source chaudes hydrothermales à plus de 100 °C, milieux hypersalins). Woese leur donna le nom de « bactéries anciennes », en imaginant que ces bactéries étaient des descendantes des premières formes de vie qui auraient existé sur la terre et qui seraient aujourd'hui réfugiées dans les quelques niches ressemblant encore à ce qu'avait pu être l'atmosphère de la terre primitive. On sait depuis peu [31] que ces bactéries sont en fait beaucoup plus largement répandues et l'inventaire systématique du milieu océanique en a révélé une remarquable diversité dans cet environnement aérobie. Il est frappant de constater que le groupe des archébactéries, défini au départ sur la base des seules données d'ARNr, a révélé ultérieurement d'autres caractéristiques biochimiques communes et en particulier une structure tout à fait originale des lipides membranaires dont les liaisons sont de type éther alors qu'elles sont de type ester pour tous les autres organismes. Quel que soit leur degré réel d'ancienneté, discuté au paragraphe suivant, il ne fait plus de doute qu'un nouveau groupe d'organismes très distant des procaryotes plus « classiques » a été identifié, même s'il persiste des débats sur la monophylie de l'ensemble des archébactéries, à savoir si certaines sont à rapprocher des eubactéries et d'autres des eucaryotes.

b. Les séquences des ARN ribosomiques des mitochondries et des plastes (voir le maïs sur la *figure 6*) les placent parmi les eubactéries, très loin de la séquence de l'ARNr nucléaire des espèces correspondantes. Ceci a mis un terme au débat sur l'origine de ces organites : origine autogène, à partir de séquences nucléaires « exportées » dans le cytoplasme, *versus* origine endosymbiotique, à partir de bactéries qui auraient pénétré dans un eucaryote.

La théorie endosymbiotique l'a emporté parce qu'il est beaucoup plus parcimonieux d'expliquer les dizaines de similitude moléculaire entre séquences mitochondriales, chloroplastiques et eubactériennes par la parenté directe entre celles-ci, plutôt que par des dizaines de convergences. La question du nombre d'endosymbioses et du moment dans l'évolution des eucaryotes auquel elles se sont produites continue à faire aujourd'hui l'objet de recherches actives [32, 33]. Une découverte récente concerne un nouveau type d'endosymbiose qui semble s'être produite au cours de l'évolution d'un groupe d'organismes photosynthétiques, les Cryptophytes (et peut-être aussi des Chromophytes) et qui met en jeu l'endosymbiose d'un eucaryote photosynthétique dans un autre eucaryote [32]. Ces deux points seront repris dans la partie 2 de cet article (à paraître dans *m/s n° 11, vol. 11, novembre 1995*).

c. Enfin, sur cet arbre, on remarque que les distances évolutives entre toutes les lignées d'organismes multicellulaires, plantes, animaux et champignons, sont relativement petites. Ainsi la multicellularité serait d'apparition relativement récente chez les eucaryotes, précédée d'une longue phase d'évolution sous forme unicellulaire. De plus, les distances évolutives, telles qu'on les mesure par cet indicateur qu'est l'ARN ribosomique, montrent que plantes, animaux et champignons sont bien moins distants les uns des autres que certains groupes de protistes ne le sont entre eux (cas du trypanosome et de l'euglène sur la *figure 6*).

D'après les ARNr, les eucaryotes multicellulaires représentent environ un vingtième de la diversité moléculaire de l'ensemble du vivant. Nous avons cependant tendance à considérer, subjectivement, qu'ils regroupent l'essentiel de la diversité phénotypique du vivant. Comme l'avait déjà souligné Wilson [34], il n'y a donc pas nécessairement corrélation entre divergence moléculaire et divergence phénotypique apparente.

Jusqu'à la fin des années 80, l'arbre universel du vivant était présenté de manière non racinée, comme sur la *figure 6*. En effet, pour raciner un arbre évolutif, il faut disposer d'un groupe que l'on sait, par des critères indépendants, être « externe » au groupe étudié. Par exemple, si l'on cherche à raciner un arbre d'espèces animales, il est raisonnable de se servir de la séquence homologue de plantes. On place alors la racine sur le segment qui relie les plantes aux animaux et l'ensemble de l'arbre des animaux s'en trouve ainsi polarisé.

Mais comment raciner l'arbre universel du vivant puisque, par définition, il englobe toutes les formes de vie et qu'on ne dispose donc pas de groupe extérieur ? Une idée ingénieuse, initialement proposée par Schwartz et Dayhoff [35], a été mise en œuvre par deux équipes en 1989 [36, 37].

L'enracinement de l'arbre universel du vivant et la question de l'origine des eucaryotes

Le principe de la méthode utilisée par Gogarten *et al.* d'une part [36] et Iwabe *et al.* d'autre part [37] est le suivant. Imaginons qu'un gène se soit dupliqué et ait donné naissance au gène 1 et au gène 2 avant que les trois lignées, Archaea, Bacteria et Eucarya, ne se séparent les unes des autres. N'importe quelle séquence du gène 1 est alors extérieure à toutes les séquences du gène 2. Les séquences du gène 1 permettent donc de raciner l'arbre du vivant obtenu avec les séquences du gène 2. On obtient en fait un arbre « en miroir » dont la racine est placée en un point médian entre les deux parties symétriques. En pratique, quelles conditions doivent être vérifiées pour que l'on puisse faire une telle hypothèse ? Il faut que toutes les espèces vivantes possèdent les deux copies de ce gène. Il faut aussi qu'il y ait plus de divergence entre l'une des copies du gène et l'autre (c'est-à-dire entre paralogues), qu'entre les versions différentes du même gène (c'est-à-dire entre orthologues), même si l'on compare ces versions entre Archaea, Bacteria et Eucarya. On s'attend alors

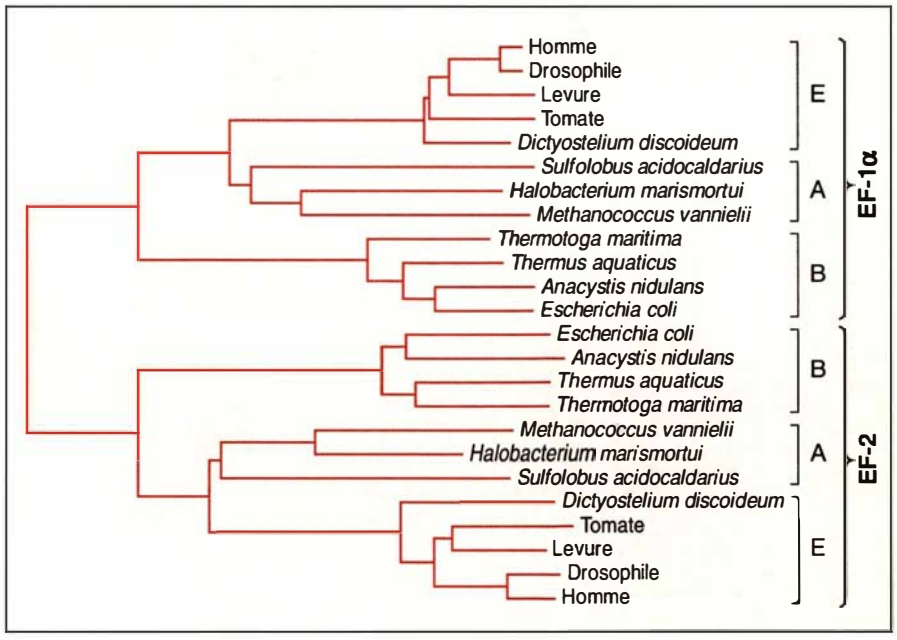

Figure 7. ***Enracinement de l'arbre du vivant.*** *Arbre construit à partir de séquences de facteurs d'élongation EF-1α et EF2 en utilisant les mêmes portions que Iwabe et al. [37]. On observe que l'arbre obtenu avec les EF-1α est presque parfaitement symétrique de l'arbre obtenu avec les EF2. Le regroupement des archébactéries avec les eucaryotes semble assez robuste puisque les valeurs de bootstrap (en %) sont de 96 pour les EF-1α, de 79 pour les EF2 et de 100 pour les ATPase F1-α et ATPase F1-β (d'après [37]).*

à ce que les deux jeux de séquences orthologues se regroupent et que l'une des jeux de séquences serve de groupe externe à l'autre et réciproquement.

Cette approche a été utilisée avec divers gènes qui semblaient remplir les conditions précisées plus haut : facteurs d'élongation (EF-1α versus EF2, appelé respectivement EF-Tu et EF-G chez les eubactéries), ATPases et déshydrogénases. Dans chacune des parties de l'arbre symétrique, désormais polarisé, on constate alors *(figure 7)* que la première lignée qui émerge est celle des Bacteria, avec pour lignée sœur, une branche qui se subdivise ultérieurement en Archaea et Eucarya. De manière quelque peu imprévue, les eucaryotes auraient pour groupe-frère les archébactéries et il n'y aurait pas de regroupement des procaryotes d'un côté de l'arbre, opposé aux eucaryotes de l'autre côté.

Un tel enracinement de l'arbre universel du vivant, rapidement accepté et introduit dans les grands manuels d'enseignement, indique que les eucaryotes dérivent de cellules ayant une organisation de type procaryote. On retrouve donc un résultat conforme aux intuitions anciennes. Ce résultat a été accepté d'autant plus facilement qu'il existe une plus grande similitude entre certaines protéines eucaryotes et archébactériennes qu'avec les mêmes protéines eubactériennes [38, 39]. Ceci est particulièrement frappant pour l'ARN polymérase : non seulement la séquence primaire de la plus grosse des sous-unités de l'enzyme archébactérienne présente une importante similitude avec celle de l'ARN polymérase II et III des eucaryotes mais la complexité globale de l'enzyme des archébactéries, avec plus d'une dizaine de sous-unités, la rapproche beaucoup plus de celle des eucaryotes que de celle des eubactéries. Il a été récemment établi que plusieurs de ces sous-unités sont de véritables homologues des sous-unités eucaryotes et, enfin, les archébactéries utilisent les mêmes facteurs de transcription TF II B et TF II D que les eucaryotes pour reconnaître la boîte TATA [40, 41].

Plusieurs analyses récentes amènent cependant à remettre en question la conclusion qu'Archaea et Eucarya sont des groupes frères. Un résumé très clair des arguments incitant à la prudence a été présenté par Forterre *et al.* [42]. En premier lieu, les relations supposées d'orthologie entre les gènes dupliqués utilisés pour raient être à revoir ; deux duplications successives pourraient avoir eu lieu pour les ATPases, ce qui conduirait à positionner la racine de manière erronée. En second lieu, ces gènes dupliqués sont séparés depuis si longtemps que leurs séquences ne sont alignables que sur de courts segments, mais surtout, que de très nombreuses substitutions multiples ont pu se produire. Or, comme nous l'avons expliqué précédemment, celles-ci peuvent complètement brouiller le message phylogénétique et les arbres obtenus sont donc peu fiables. Pour les gènes EF1α et EF2, la *figure 8* montre que les séquences sont saturées en substitutions, laissant à penser que, comme pour la thésaurine *(figure 3)*, la position basale des eubactéries pourraient être due au phénomène d'attraction des longues branches. Enfin, si plusieurs gènes montrent effectivement plus de similitudes entre archébactéries et eucaryotes, toutes les autres configurations sont obtenues et, en particulier, le nombre de gènes d'eucaryotes s'apparentant plutôt à ceux des eubactéries est loin d'être négligeable. Parmi ces gènes, il faut néanmoins distinguer ceux qui ont pour origine les symbiontes bactériens, destinés à devenir respectivement les mitochondries et les plastes, d'autant plus que ceux-ci ont massivement transféré leurs gènes vers le noyau de l'hôte. Cette catégorie ne nous éclaire pas sur l'origine des gènes du « proto-eucaryote ».

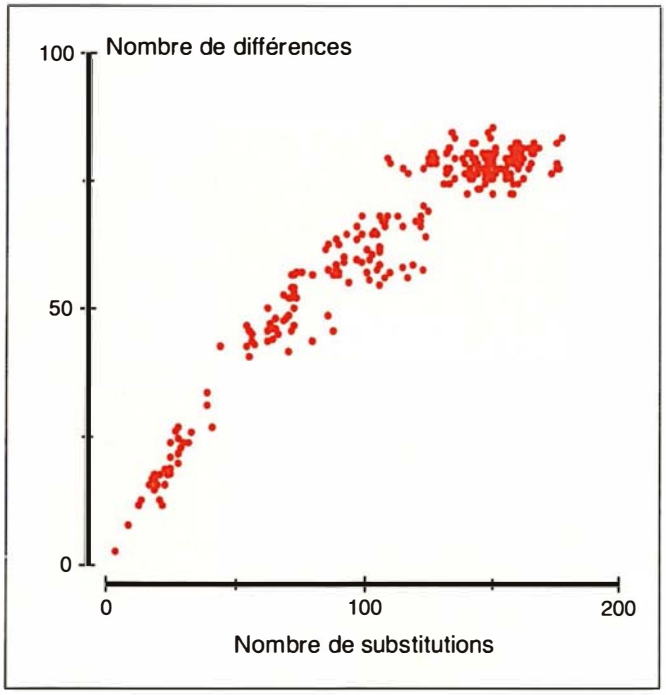

Figure 8. ***Les archébactéries sont-elles proches des eucaryotes ?*** *Mise en évidence de la saturation avec la méthode utilisée pour la figure 3A appliquée aux séquences ayant servi à construire l'arbre de la figure 7. On constate que le nombre de différences plafonne à 75 alors que le nombre de substitutions atteint 175. Une si grande fréquence des substitutions multiples amène à remettre en cause la proche parenté entre archébactéries et eucaryotes suggérée par ces gènes, ce qui est en accord avec d'autres marqueurs phylogénétiques [42].*

Ainsi, certaines portions du génome eucaryote semblent s'apparenter à celui des archébactéries, et d'autres à celui des eubactéries. Ceci a suggéré que le génome eucaryote pourrait être en fait une mosaïque évolutive et a conduit, sous sa forme extrême, à une hypothèse proposée par Zillig [43] et d'autres qui consiste à considérer le génome eucaryote comme résultant entièrement d'une juxtaposition de séquences d'archébactéries et d'eubactéries (avec perte de l'un ou l'autre des gènes en « double »). A titre d'hypothèse, Forterre [44], pour sa part, privilégie un troisième scénario dans lequel les eucaryotes ont une origine très ancienne, se séparant dès l'origine des procaryotes (Archaea + Bacteria). L'évolution vers la compaction du génome et vers le couplage transcription-traduction de ces deux derniers groupes étant dérivée et ayant été guidée par une pression pour l'adaptation à la vie à haute température à partir d'un ancêtre « mésophile » (vivant à température moyenne). Dans cette hypothèse, de nombreux traits de la cellule eucaryote ne seraient plus dérivés mais primitifs (introns, chromosomes linéaires, découplage transcription-traduction, etc.).

En définitive, la question de l'enracinement de l'arbre universel du vivant et de l'origine des eucaryotes reste encore ouverte [30, 42]. Il est encore bien difficile de choisir entre le modèle où les eucaryotes sont « tardifs » et groupe-frère des archébactéries, celui où ils sont « précoces » et séparés dès l'origine des procaryotes, et enfin celui où ils n'ont jamais eu d'existence ancienne propre mais où ils seraient issus d'une fusion entre une archébactérie et une eubactérie.

La paléontologie ne permet malheureusement pas de trancher cette question. Les fossiles de type procaryote (considérés comme apparentés aux cyanobactéries) sont bien les premiers à apparaître dans les couches géologiques (environ 3,5 milliards d'années), alors que des fossiles de type eucaryote n'apparaissent que beaucoup plus tard (1,5 milliards d'années mais peut-être 2,1). Mais cette absence d'eucaryotes précoces pourrait résulter à la fois de problèmes de préservation et de l'insuffisance des critères servant à leur caractérisation (principalement la taille). Notons seulement qu'en l'état actuel, les données paléontologiques sont compatibles avec un scénario d'apparition « tardive » des eucaryotes [45].

Le réticulum endoplasmique et le cytosquelette : des homologues chez les procaryotes ?

Quelle que soit la position de la racine de l'arbre universel du vivant, on peut se demander parmi les caractéristiques de chacun des trois grands groupes, Archaea, Bacteria et Eucarya, quelles sont celles qui ont été héritées de l'ancêtre commun. Par exemple, les caractéristiques de la cellule eucaryote (noyau, compartimentation membranaire intracellulaire, cytosquelette) ont le plus souvent été interprétées comme des acquisitions de la lignée eucaryote, à partir d'un ancêtre de type procaryote. Des résultats récents amènent cependant à reconsidérer ce point.

Bactéries comme Eucaryotes doivent disposer de systèmes complexes permettant la translocation des protéines au travers des membranes. Chez les eucaryotes, la fonction est principalement assurée par le « translocateur » du reticulum endoplasmique qui permet le passage dans la lumière du reticulum (ou l'intégration dans sa membrane) des protéines sécrétoires, golgiennes, lysosomiales et de celles destinées à la membrane plasmique. Chez les bactéries, les protéines se dirigent du cytoplasme vers l'espace périplasmique, la membrane externe (chez les gram$^-$) ou le milieu extérieur. Topologiquement, la ségrégation des protéines dans la lumière du reticulum chez les eucaryotes s'apparente à leur exclusion du cytosol et à leur transfert vers un compartiment qui est l'équivalent de l'extérieur de la cellule. Il n'y a en somme qu'une étape supplémentaire par rapport au transfert direct vers le milieu extracellulaire qui a lieu chez les bactéries [2].

Partant de ces notions, Blobel [46] a proposé voici plusieurs années que le reticulum endoplasmique ait pris naissance comme une invagination de la membrane plasmique d'une bactérie, qui se serait ensuite autonomisée. Les translocateurs servant à l'exportation chez les bactéries se retrouveraient ainsi dans la membrane du reticulum, et, sans inversion d'orientation, dirigeraient les protéines vers la lumière de ce dernier. L'enveloppe nucléaire qui est en continuité avec le reticulum par sa membrane externe aurait pu être générée au cours de ce processus ; on explique donc par un seul événement deux caractéristiques-clés de la cellule eucaryote. S'il en a bien été ainsi, les composants du translocateur bactérien et de celui du reticulum auraient une origine évolutive commune et leur parenté pourrait être encore détectable. Depuis la fin des années 70, tant le système bactérien que celui du reticulum ont été soumis à une dissection génétique et biochimique intensive. Ceci a permis l'identification des principaux composants protéiques de chacun des systèmes, l'identification des gènes correspondants et, depuis peu, la séquence de plusieurs de ces gènes. La prédiction s'est ainsi trouvée vérifiée par la mise en évidence d'homologies entre certaines des protéines du translocateur bactérien et celles du translocateur eucaryote [47].

Les translocateurs bactériens et eucaryotes ont donc une origine évolutive commune. Bien que ceci soit en accord avec l'hypothèse que le reticulum ait pour origine la membrane plasmique d'un procaryote, notons cependant que cela n'en fournit pas une preuve absolue. Les eucaryotes n'ont eu qu'assez peu de « bricolage » à réaliser pour mettre au point l'un de leurs compartiments les plus caractéristiques, le vaste réseau de membranes internes incluant l'enveloppe nucléaire.

Un autre trait frappant de tous les eucaryotes, dont on a longtemps cru qu'il n'avait aucun correspondant chez les procaryotes, est l'existence d'une gamme d'éléments fibrillaires collectivement dénommés cytosquelette. Microtubules (principalement constitués de tubuline) et microfilaments (principalement constitués

d'actine) sont en particulier des éléments totalement ubiquitaires chez les eucaryotes, y compris dans les rameaux les plus anciennement séparés de l'ancêtre commun de tous les eucaryotes. Pour réaliser leur fonction d'organisation de l'espace intracellulaire et de motilité, ces fibrilles cytosquelettiques s'associent avec de volumineuses ATPases qui jouent le rôle de « moteurs » (dynéines et kinésines des microtubules, myosines des microfilaments, etc.).

Un résultat surprenant de ces dernières années est la mise en évidence chez certaines bactéries de protéines qui pourraient avoir une parenté évolutive avec celles du cytosquelette des eucaryotes. Le cas le plus fascinant est celui de la protéine Fts Z de *E. coli*. Initialement identifiée par une mutation conduisant à un défaut de septation, on a ensuite montré que la protéine se localisait en anneau à l'apex du septum en cours de croissance [48], puis qu'il s'agissait d'une GTPase (comme la tubuline) dont une très courte séquence de 7 acides aminés était identique, à un acide aminé près, à une séquence hautement conservée dans toutes les tubulines [49]. Enfin, on vient d'établir que la protéine Fts Z, surproduite par génie génétique, est susceptible de s'assembler *in vitro* en tubules (de diamètre différent des microtubules) et que cet assemblage, comme celui de la tubuline, ne se faisait qu'en présence de GTP [50]. L'ensemble de ces similitudes (y compris leur implication dans la division cellulaire) est frappant ; il reste à vérifier qu'elles correspondent à de véritables homologies évolutives.

Citons aussi le gène muk B de *E. coli* dont les mutations conduisent à des défauts de ségrégation de l'ADN bactérien au cours de la division, et qui code une grande protéine avec une tête susceptible de lier un nucléotide triphosphate et une longue queue susceptible de former des hélices-α torsadées (*coiled-coil*). Bien qu'aucune similitude n'ait été détectée avec les séquences de myosine d'eucaryote, cette protéine en a plusieurs caractéristiques et elle semble également impliquée dans une fonction motrice, à la seule différence qu'ici elle met en jeu la ségrégation de l'ADN [51].

En conclusion, si les eucaryotes sont d'apparition plus tardive que les procaryotes, plusieurs des traits qui nous paraissent les différencier si profondément des procaryotes pourraient en fait avoir des homologues chez ces derniers. Les eucaryotes n'auraient eu qu'à les développer et les spécialiser, par exemple, en ayant recours aux duplications et réarrangements de gènes qui sont un autre trait si caractéristique des eucaryotes. Bien que les relations de parenté des eucaryotes avec les eubactéries et les archébactéries ne soient pas totalement résolues, il faut souligner que la monophylie des eucaryotes est très hautement vraisemblable. Ainsi, même si la phylogénie moléculaire n'a pas encore permis de définir complètement la structure de l'arbre universel du vivant, du moins a-t-elle permis d'inférer une vue beaucoup plus équilibrée de la diversité des eucaryotes par rapport à celle des procaryotes, de subdiviser ces derniers en grands groupes très distants et, enfin, de démontrer l'origine endosymbiotique des organites cytoplasmiques à ADN. Dans le second article, nous concentrerons notre analyse sur les eucaryotes et, au travers de l'arbre phylogénétique détaillé que nous présenterons, nous discuterons de la nature des premiers eucaryotes, des stades de réalisation des symbioses mitochondriales et chloroplastiques, et enfin de l'origine des organismes multicellulaires ■

RÉFÉRENCES

1. Haeckel E. *Anthropogenie: Keimes – und Stammes – Geschichte des Menschen.* Leipzig : W. Engelmann, 1874 ; 732 p.

2. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. *Molecular Biology of the Cell.* New York, London : Garland Publishing Inc., 1983 and 1989.

3. Sédillot CE. De l'influence des découvertes de Monsieur Pasteur sur les progrès de la chirurgie. *CR Acad Sci Paris* 1878 ; 86 : 634.

4. Chatton E. Titres et travaux scientifiques. Sottano, Italy : Sete, 1937.

5. Stanier RY, van Niel CB. The concept of a bacterium. *Arch Microbiol* 1942 ; 42 : 17-35.

6. Hennig W. Phylogenetic systematics. Urbana: University of Illinois Press, 1966.

7. Whittaker RH. On the broad classification of organisms. *Q Rev Biol* 1959 ; 34 : 210-26.

8. Swofford DL, Olsen GJ. Phylogeny reconstruction. *In:* Hillis DM, Moritz C, eds. *Molecular Systematics* 1990 ; 411-501.

9. Li WH, Graur D. Fundamentals of molecular evolution. Sunderland Mass: Sinauer Associates Inc. 1991.

10. Darlu P, Tassy P. Reconstruction phylogénétique. Paris: Masson, 1993 ; 245 pp.

11. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull* 1958 ; 38 : 1409-38.

12. Li WH. So, what about the molecular clock hypothesis? *Curr Op Gen Dev* 1993 ; 3 : 896-901.

13. Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 1978 ; 27 : 401-10.

14. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985 ; 39 : 783-91.

15. Hendy MD, Penny D. A framework for the quantitative study of evolutionary trees. *Syst Zool* 1989 ; 38 : 297-309.

16. Lecointre G, Philippe H, Lê HLV, Le Guyader H. Species sampling has a major impact on phylogenetic inference. *Mol Phylo Evol* 1993 ; 2 : 205-24.

17. Philippe H, Douzery E. The pitfalls of molecular phylogeny based on four species as illustrated by the Cetacea/Artiodactyla relationships. *J Mam Evol* 1994 ; 2 : 133-152.

18. Zharkikh A, Li WH. Inconsistency of the maximum-parsimony method : the case of five taxa with a molecular clock. *Syst Biol* 1993 ; 42 : 113-25.

19. Takezaki N, Nei M. Inconsistency of the maximum-parsimony method when the rate of nucleotide substitution is constant. *J Mol Evol* 1994 ; 39 : 210-8.

20. Fitch WM, Markowitz E. An improved method for determining codon variability in a gene and its application to the rate of fixations of mutations in evolution. *Biochem Genet* 1970 ; 4 : 579-93.

21. Palumbi SR. Rates of molecular evolution and the fraction of nucleotide position free to vary. *J Mol Evol* 1989 ; 29 : 180-7.

22. Viel A, Lemaire M, Philippe H, MoralesJ, Mazabraud A, Denis H. Structural and functionnal properties of thesaurin a (42Sp50), the major protein of the 42S particles present in *Xenopus laevis* previtellogenic oocytes. *J Biol Chem* 1991 ; 266 : 10392-9.

RÉFÉRENCES

23. Baldauf SL, Palmer JD. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci USA* 1993 ; 90 : 11558-62.

24. Lecointre G, Philippe H, Lê HLV, Le Guyader H. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. *Mol Phylo Evol* 1994 ; 3 : 292-309.

25. Philippe H, Chenuil A, Adoutte A. Can the cambrian explosion be inferred through molecular phylogeny ? *Development* 1994 ; 120 (suppl) 15-25.

26. Philippe H, Adoutte A. What can phylogenetic patterns tell us about the evolutionary processes generating biodiversity? In: Hochberg M, Clobert J, Barbault R, eds. *Aspects of the genesis and maintenance of biological diversity*. Oxford: Oxford University Press, 1995 (sous presse).

27. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* 1977 ; 74 : 5088-90.

28. Olsen GJ. Earliest phylogenetic branchings : comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symp Quant Biol* 1987 ; 52 : 825-37.

29. Olsen GJ, Woese CR. Ribosomal RNA : a key to phylogeny. *FASEB J* 1993 ; 113-23.

30. Doolittle WF, Brown JR. Tempo, mode, the progenote, and the universal root. *Proc Natl Acad Sci USA* 1994 ; 91 : 6721-8.

31. De Long EF, Wu KY, Prézelin BB, Jovine RVM. High abundance of Archaea in Antarctic marine picoplankton. *Nature* 1994 ; 371 : 695-7.

32. Gray MW. The endosymbiont hypothesis revisited. In : Wolstenholme DR, Jeon KW, eds. *Mitochondrial genomes*. San Diego, California : Academic Press Inc, 1992.

33. Loiseaux-de Goër S. Plastid lineages. In : Round FE, Chapman DJ, eds. *Progress in phycological research, vol. 10*. Biopress Ltd, 1994 ; 137-177.

34. Wilson AC, Maxson LR, Sarich VM. Two types of molecular evolution. Evidence from studies of interspecific hybridization. *Proc Natl Acad Sci USA* 1974 ; 71 : 2843-7.

35. Schwartz RM, Dayhoff MO. Chloroplast origins : inferences from protein and nucleic acid sequences. *Science* 1978 ; 199: 395-403.

36. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi J, Denda K, Yoshida M. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc Natl Acad Sci USA* 1989 ; 86: 6661-5.

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

38. Zillig W, Schnabel R, Stetter KO. Archaebacteria and the origin of the eukaryotic cytoplasm. *Curr Top Microbiol Immunol* 1985 ; 114 : 1-18.

39. Huet J, Schnabel R, Sentenac A, Zillig W. Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type. *EMBO J* 1983 ; 2 : 1291-4.

40. Keeling PJ, Charlebois RL, Doolittle WF. Archaebacterial genomes : eubacterial form and eukaryotic content. *Curr Op Gen Dev* 1994 : 4, 816-22

41. Klenk H-P, Doolittle WF. Archaea and eukaryotes versus bacteria ? *Curr. Biol* 1994 ; 4: 920-2.

42. Forterre P, Benachenhou-Lahfa N, Confalonieri F, Duguet M, Elie C, Labedan B. The nature of the last universal ancestor and the root of the tree of life, still open questions. *BioSystems* 1993 ; 28 : 15-32.

43. Zillig W, Klenk HP, Palm P, Leffers H., Pühler G, Gropp F, Garrett R. Did eukaryotes originate by a fusion event? *Endocytobiosis Cell Res* 1989 ; 6 : 1-25.

44. Forterre P. Thermoreduction, a hypothesis for the origin of prokaryotes. *CR Acad Sci Paris* 1995 ; 318 : 415-22.

45. Knoll AH. The early evolution of eukaryotes: A geological perspective. *Science* 1992 ; 256 : 621-7.

46. Blobel G. Intracellular protein topogenesis. *Proc Natl Acad Sci USA* 1980 ; 77 : 1496-500.

47. Hartmann E., Sommer T, Prehn S, Görlich D, Jentsch S, Rapoport TA. Evolutionary conservation of components of the protein translocation complex. *Nature* 1994 ; 367 : 654-7.

48. Bi E, Lutkenhaus J. FtsZ ring structure associated with division in *Escherichia coli*. *Nature* 1991 ; 354 : 161-4.

49. de Boer P, Crossley R, Rothfield L. The essential bacterial cell-division protein FtsZ is a GTPase. *Nature* 1992 ; 359 : 254-6.

50. Bramhill D, Thompson CM. GTP-dependent polymerization of *Escherichia coli* FtsZ protein to form tubules. *Proc Natl Acad Sci USA* 1994 ; 91 : 5813-7.

51. Niki H, Jaffé A, Imamura R, Ogura T, Hiraga S. The new gene *mukB* codes for a 177 kd protein with coiled-coil domains involved in chromosome partitioning of *E. coli*. *EMBO J* 1991 ; 10 : 183-93.

52. Philippe H, Sõrhannus U, Baroin A, Perasso R, Gasse F, Adoutte A. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. *J Evol Biol* 1994 ; 7 : 247-65.

André Adoutte
Hervé Philippe
Hervé Le Guyader

Laboratoire de biologie cellulaire, 4, URA 1134 Cnrs, bâtiment 444, université Paris-Sud, 91405 Orsay Cedex, France.

Agnès Germot

Laboratoire de biologie comparée des protistes, URA 1944 Cnrs, université de Clermont-Ferrand 2, Les Cézeaux, 24, avenue des Landais, 63177 Aubière Cedex, France.

Remerciements

Nous remercions la SFG et, en particulier, Monique Fuquhara et Alain Nicolas pour leurs incitations à rédiger cet article et plusieurs collègues, tout particulièrement Jean Génermont, pour leur lecture critique du texte.

TIRÉS À PART

A. Adoutte.