Université de Montréal

Phylogénomique des Archées

par

Jean-Christophe Grenier

Programme de Bio-informatique, Département de Biochimie

Faculté de Médecine

Mémoire présenté à la Faculté de Médecine

en vue de l'obtention du grade de M.Sc.

en Bio-Informatique

Juillet 2011

© Jean-Christophe Grenier, 2011

Université de Montréal

Faculté des études supérieures et postdoctorales

Ce mémoire (ou cette thèse) intitulé(e) :

Phylogénomique des Archées

présenté(e) par :

Jean-Christophe Grenier

a été évalué(e) par un jury composé des personnes suivantes :

Président-rapporteur : Miklós Csuros

Directeur de recherche : Hervé Philippe

Membre du jury : Simon Joly

Résumé

Les transferts horizontaux de gènes (THG) ont été démontrés pour jouer un rôle important dans l'évolution des procaryotes. Leur impact a été le sujet de débats intenses, ceux-ci allant même jusqu'à l'abandon de l'arbre des espèces. Selon certaines études, un signal historique dominant est présent chez les procaryotes, puisque les transmissions horizontales stables et fonctionnelles semblent beaucoup plus rares que les transmissions verticales (des dizaines contre des milliards). Cependant, l'effet cumulatif des THG est non-négligeable et peut potentiellement affecter l'inférence phylogénétique. Conséquemment, la plupart des chercheurs basent leurs inférences phylogénétiques sur un faible nombre de gènes rarement transférés, comme les protéines ribosomales. Ceux-ci n'accordent cependant pas autant d'importance au modèle d'évolution utilisé, même s'il a été démontré que celui-ci est important lorsqu'il est question de résoudre certaines divergences entre ancêtres d'espèces, comme pour les animaux par exemple.

Dans ce mémoire, nous avons utilisé des simulations et analyser des jeux de données d'Archées afin d'étudier l'impact relatif des THG ainsi que l'impact des modèles d'évolution sur la précision phylogénétique. Nos simulations prouvent que (1) les THG ont un impact limité sur les phylogénies, considérant un taux de transferts réaliste et que (2) l'approche super-matrice est plus précise que l'approche super-arbre. Nous avons également observé que les modèles complexes expliquent non seulement mieux les données que les modèles standards, mais peuvent avoir un impact direct sur différents groupes phylogénétiques et sur la robustesse de l'arbre obtenu. Nos résultats contredisent une publication récente proposant que les Thaumarchaeota apparaissent à la base de l'arbre des Archées.

Mots-clés : phylogénie, phylogénomique, procaryotes, Archées, transfert horizontal de gènes, évolution moléculaire, simulations, modèles évolutifs, super-matrice, super-arbre.

Abstract

Horizontal gene transfer (HGT) had been demonstrated to play an important role in the evolution of prokaryotes. Their impact on phylogeny was the subject of a heated debate, with some proposing that the concept of a species tree should be abandoned. The phylogeny of prokaryotes does contain a major part of the historical signal, because stable and functional horizontal transmissions appear to be by far rarer than vertical transmissions (tens versus billions). However, the cumulative effect of HGT is non-negligible and can potentially affect phylogenetic inference. Therefore, most researchers base their phylogenetic inference on a low number of rarely transferred genes such as ribosomal proteins, but they assume the selection of the model of evolution as less important, this despite the fact that it has been shown of prime importance for much less deep divergences, e.g. like animals.

Here, we used a combination of simulations and of real data from Archaea to study the relative impact of HGT and of the inference methods on the phylogenetic accuracy. Our simulations prove that (1) HGTs have a limited impact on phylogeny, assuming a realistic rate and (2) the supermatrix is much more accurate than the supertree approach. We also observed that more complex models of evolution not only have a better fit to the data, but can also have a direct impact on different phylogenetic groups and on the robustness of the tree. Our results are in contradiction to a recent publication proposing that the Thaumarchaeota are at the base of the Archaeal tree.

Keywords : phylogeny, phylogenomics, prokaryotes, Archaea, horizontal gene transfer, molecular evolution, simulation, evolutionary models, supermatrix, supertree.

Table des matières

Liste des tableaux

Liste des figures

Sigles et abréviations

A	Adénine
C	Cytosine
G	Guanine
T	Thymine
R	Purine (A et G)
Y	Pyrimidine (C et T)
ADN	Acide désoxyribonucléique
ALB/LBA	Attraction des longues branches
ARN	Acide ribonucléique
ARNt	ARN de transfert
ARNr	ARN ribosomique
BV	Valeur de Bootstrap (Support statistique)
GTR	« *General Time Reversible* »
kb	Kilobase
LUCA	Dernier ancêtre commun universel "*Last universal common ancestor*"
MAF	« Maximum agreement forest »
nm	Nanomètre
NNI	« *Nearest Neighbor Interchange* »
PP	Probabilités postérieures (Support statistique en inférence Bayésienne)
SPR	« Subtree pruning and regrafting »
SSU	Petite sous-unité (« *small subunit* »)
PCA	Analyse de composantes principales (« Principal component analysis »)
THG/HGT	Transfert horizontal de gène
ToL	Arbre de la vie (« Tree of Life »)

Dédicace

À Gabrielle, ma bien aimée.

Remerciements

Je tiens à remercier tout particulièrement mon directeur **Hervé Philippe**, avec qui j'ai eu le plaisir de travailler durant mes études à la maîtrise. **Hervé** possède à mon avis toutes les qualités d'un grand chercheur, soit par sa patience, la justesse de ses analyses, la rigueur de son travail ainsi que l'objectivité avec laquelle il traite ses projets. Je tiens également à remercier **Henner Brinkmann**, qui est toujours présent pour répondre à nos questions et qui a grandement collaboré à l'écriture des manuscrits ainsi qu'à la conception de plusieurs des concepts expérimentaux. Je voudrais également remercier les membres présents et passés du laboratoire qui ont su collaborer de près ou de loin aux différentes étapes de mon cheminement académique. Un grand merci à **Claudia Kleinman** pour ces nombreux conseils, à **Béatrice Roure, Steven Hébert** ainsi qu'à **Nicolas Lartillot**, pour différentes suggestions et l'utilisation de ses ressources informatiques, et à **Raphaël Poujol**. Un merci tout spécial également à **Marie-Ka Tilak** qui a su égailler mes journées pendant près d'un an et demi avec sa bonne humeur et les discussions constructives auxquelles nous avons eu droit. Merci aussi à **Simon Laurin-Lemay** pour la relecture des manuscrits et sa perception biologique du travail accompli. Je remercie également mon parrain de maîtrise, **François-Joseph Lapointe** pour les quelques rencontres que nous avons eues concernant mon projet.

Un merci tout particulier également au personnel du Département de Biochimie de l'Université de Montréal dont **Gertraud Burger** pour avoir instauré le programme de bourse biT qui m'a permis de survivre pendant un an et demi ainsi qu'à **Élaine Meunier**, sans qui nous serions sans cesse en retard pour la paperasse.

Un merci tout spécial finalement aux membres de ma famille et plus particulièrement à mes parents, **Jocelyn** et **France**, qui m'ont toujours encouragé dans mes démarches. Merci également à **Gabrielle**, ma copine, de m'avoir supporté et encouragé ces dernières années.

Chapitre 1

Introduction - Revue de la littérature

1.1 Arbre universel de la vie

1.1.1 L'arbre des espèces

De nos jours, il est majoritairement reconnu que toute espèce vivante sur cette planète descend d'un même ancêtre commun, aussi appelé LUCA (*Last Universal Common Ancestor*) (Forterre et Philippe, 1999). Nous reconnaissons cela puisque chaque espèce utilise le même code génétique (ou de très légers variants) et possèdent une machinerie cellulaire semblable. Le concept d'une origine commune à toutes les espèces a été proposé par Charles Darwin en 1859 dans son célèbre ouvrage *On the origin of species by means of natural selection*. La théorie de Darwin comporte trois concepts principaux : il existe une grande variabilité à l'intérieur de chaque population, la sélection naturelle, c'est-à-dire que les formes les mieux adaptées à l'environnement survivront, cause la descendance avec modifications au cours des générations et le système naturel de classification est généalogique (Dayrat, 2003). Les relations de parentés entre les espèces y étaient proposées comme étant sous la forme d'un arbre, contenant une racine, pour le dernier ancêtre commun, les nœuds pour les ancêtres communs de différents groupes ainsi que les feuilles, représentant les dernières espèces de ces lignées. Si la feuille se situe à un point inférieur au temps actuel, il s'agit d'une espèce éteinte (figure 1.1). Les relations de parenté entre les différentes espèces, ou l'histoire évolutive des espèces, ont pour la première fois été représentées sous forme d'arbre par Ernest Haeckel (1866). C'est d'ailleurs Haeckel qui créa le mot « phylogénie » afin de décrire ces relations.

Cependant, au cours des dernières années, plusieurs mécanismes autres que la spéciation ont été découverts. Les phénomènes de l'hybridation, l'introgression, l'endosymbiose ou les transferts horizontaux de gènes (THG), ont remis en question la vision de l'évolution qu'avait Darwin, le monisme (un seul mode d'évolution), et celle-ci

s'est heurtée à plusieurs autres théories. Le pluralisme est l'une de celles-ci et évoque la possibilité que plusieurs modes différents d'évolution puissent exister pour différents taxons, ou à plus grande échelle taxonomique (Doolittle et Bapteste, 2007). En d'autres termes, un arbre ne pourrait représenter totalement l'ensemble de l'histoire évolutive des organismes. Les arguments sont surtout dirigés vers la difficulté d'application de ce modèle unique pour les organismes microbiens, qui utilisent des méthodes de transmission non-verticales de leur matériel génétique. Cette remise en question est l'un des sujets les plus chaudement discutés présentement dans l'évolution des procaryotes et va même jusqu'à reconsidérer la notion d'arbre du vivant, un sujet qui sera abordé lors du chapitre 3 de ce mémoire.

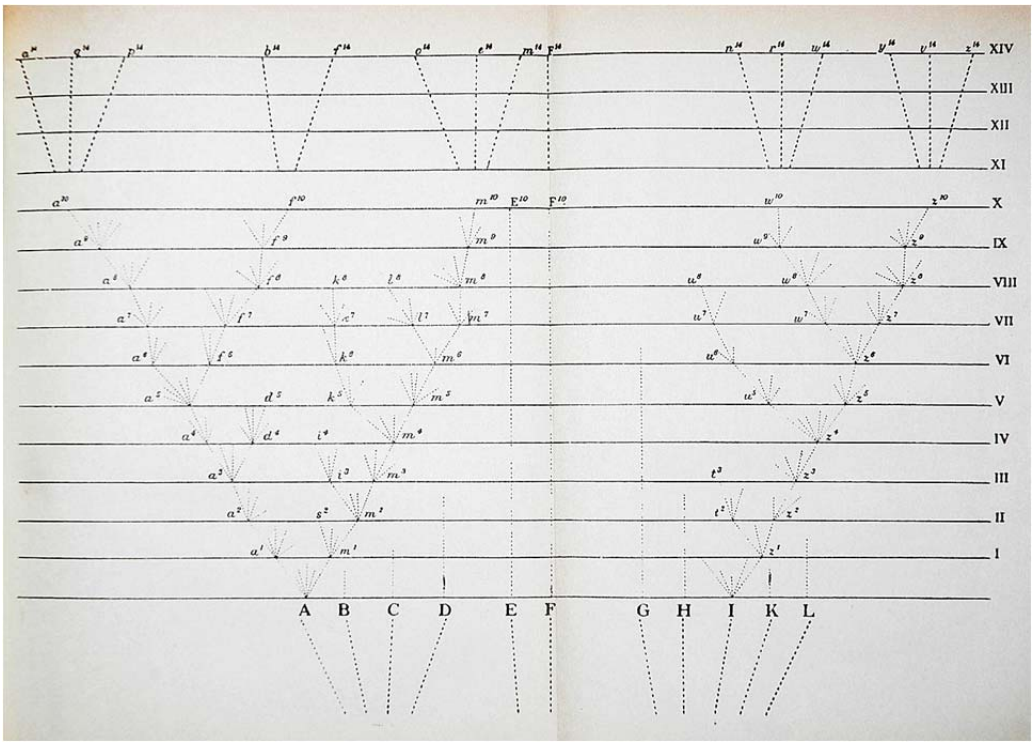


Figure 1.1 Représentation de la divergence des espèces au cours du temps (tiré de (Darwin, 1859)). L'axe verticale représente différents points dans le temps (où XIV représente l'état actuel des lignées évolutives).

1.1.2 **Les trois domaines du vivant**

En 1866, la première forme de classification du vivant telle que nous la connaissons aujourd'hui a été publiée. Ernst Haeckel propose la classification des différentes formes de vie connues à cette époque et il formera trois groupes majeurs, ou

royaumes : Plantae, Protista et Animalia (Dayrat, 2003). Ses arbres, et autres arbres primitifs, étaient fondés sur une idée générale d'une hiérarchie de relations entre les espèces et les différents niveaux taxonomiques. Avec le temps, des critères quantitatifs et objectifs sont apparus afin de définir les degrés de différence entre les espèces (Fitch et Margoliash, 1967; Wolf et al., 2002).

L'utilisation de critères de plus en plus précis et nombreux a mené à des classifications différentes des organismes. La découverte de l'existence de différents types de cellule, soient les cellules avec ou sans noyau et possédant ou ne possédant pas d'organites, a mené à la classification des espèces en deux domaines : les procaryotes et les eucaryotes. Ce n'est que relativement récemment que Woese, à l'aide d'analyses de l'ARN ribosomique 16S/18S, proposa trois domaines de la vie (Woese et Fox, 1977). Ce qui était auparavant les procaryotes, et qui rassemblait tous les organismes unicellulaires sans noyau, fut subdivisé en deux. Les trois domaines sont alors les Eucaryotes, les Eubactéries et les Archaebactéries, et font partie de l'actuel modèle standard de la phylogénie (Woese, 1987; Doolittle et Handy, 1998). Leur nom a été révisé par la suite, il s'agit désormais des Eucaryotes, des Bactéries et des Archées (Woese, Kandler et Wheelis, 1990).

La structure de l'arbre de la vie représentant les trois domaines du vivant est l'un des sujets encore chaudement débattus en évolution. L'hypothèse la plus généralement acceptée suppose que les Archées et les Eucaryotes sont groupes frères, tandis que les Bactéries dérivent directement de LUCA, le dernier ancêtre commun. Ceci fut proposé grâce à une phylogénie effectuée à partir de gènes anciennement dupliqués, les facteurs d'élongation et les ATPase (Gogarten et al., 1989; Iwabe et al., 1989). Plusieurs scénarios évolutifs ont été proposés afin d'expliquer cette topologie. Le premier consisterait à ce que les Archées et les Eucaryotes partagent un même ancêtre commun. Un autre scénario, qui est de plus en plus populaire, propose que les Archées et les Bactéries proviennent directement de LUCA et que les Eucaryotes proviennent d'une fusion entre deux membres des deux autres groupes. Cette fusion expliquerait en partie la nature chimérique des Eucaryotes (Martin et Müller, 1998; Lopez-Garcia et Moreira, 1999). Un dernier scénario propose que LUCA était de type bactérien et que les Archées

et les Eucaryotes soient dérivés d'une lignée particulière de bactérie Gram-positive (Cavalier-Smith, 2002). Cette hypothèse diffère des autres puisque les bactéries seraient dans ce cas paraphylétiques.

1.1.3 **Méthodes de classification du vivant**

Jusque vers le milieu du XX^{e} siècle, les méthodes utilisées par les morphologistes avaient de profondes lacunes et manquaient de rigueur, ce qui rendait la répétition de leurs manipulations d'identification de caractères presque impossible (surtout sur les plantes et les animaux). Il n'existait pas de méthodes assez rigoureuses pour **i)** définir les caractères (surtout leurs états) et afin de **ii)** choisir le meilleur arbre évolutif. Ceux-ci étaient même considérés plus comme des artistes que des scientifiques. Leurs critères de reconnaissance des phénotypes étaient subjectifs et sont rapidement devenus inefficaces concernant les procaryotes. Afin de résoudre le point **i)** et afin d'éviter toute subjectivité, plusieurs chercheurs proposèrent vers la fin des années '50 de nouveaux moyens, plus objectifs et plus quantitatifs. Ces moyens, disaient-ils, cherchaient à rapprocher le domaine de la vraie science et visaient à mathématiser les méthodes dans le but d'en augmenter la reproductibilité et de les rendre plus empiriques (Hull, 1985). Deux grandes méthodes de classification ont été utilisées dans le but d'identifier le meilleur arbre **(ii)** et celles-ci s'opposaient sur le concept de base sur lequel ces deux méthodes reposent. Il y a les méthodes cladistiques (Hennig, 1966) qui font reposer la classification sur les états dérivés et partagés entre les espèces et font une distinction entre caractères primitifs ou évolués, donnant un plus grand poids aux caractères évolués. Quant-à elles, les méthodes phénétiques se basent sur le nombre de caractères similaires entre les différentes espèces pour ainsi estimer une distance évolutive inter-espèce (Jensen, 2009). Un grand nombre de caractères est nécessaire pour cette méthode, puisqu'elle ne fait pas de distinction entre plésiomorphies, état ancestral d'un groupe, et apomorphies, traits de caractère différents de l'état ancestral (Hull, 1980). Un faible nombre de ces caractères peut poser problème pour l'établissement des liens de parenté advenant le cas où deux traits de caractère identiques soient apparus indépendamment chez deux espèces éloignées, caractères analogues (Williams et Ebach, 2009).

Ce n'est qu'en 1965 (Zuckerkandl et Pauling, 1965) que les caractères moléculaires furent utilisés pour la première fois et que l'impact des ces méthodes sur la classification du vivant fut observée d'une façon plus concrète. Plusieurs auteurs comme Fitch *et al.* (1967) suivirent cette tendance et initièrent pour de bon ce mouvement en proposant les gènes (protéines) comme éléments permettant de reconstruire l'histoire évolutive de la vie. L'arrivée d'un critère tel que le matériel génétique venait résoudre le problème de subjectivité lié aux états de caractères **(i)**, même s'il ne faisait toujours pas l'unanimité à l'époque. Ce n'est par contre qu'à partir de 1975 et surtout 1985 que la phylogénie moléculaire a commencé à améliorer la classification des espèces (Woese et Fox, 1977; Lane et al., 1985; Woese, 1987; Field et al., 1988). Entre autres, l'utilisation du matériel génétique comme traits de caractère est venue régler le problème d'ordre statistique qui existait pour les méthodes phénétiques puisque désormais, des milliers de caractères pouvaient être utilisés au lieu d'une dizaine. Du côté des méthodes cladistiques, la difficulté apportée par les données moléculaires concerne la recherche des caractères informatifs, qui contiennent des apomorphies. Le principe de parcimonie, qui prend en compte le scénario représentant le moins de transformations, sera appliqué dans ce cas-ci. L'utilisation du matériel génétique cause toujours quelques mésententes entre les morphologistes actuels et les phylogénéticiens quant à la classification des espèces. Ceci est causé parce que de nombreuses erreurs persistaient dans la reconstruction des arbres et ont mis du temps à être résolues, notamment à cause des différentes vitesses évolutives pour le cas des rongeurs par exemple.

Les données moléculaires ont facilité l'élaboration de modèles mathématiques afin de déterminer le niveau de similarité entre les différentes espèces et de prendre en compte différents concepts évolutifs. Ces modèles ont cependant mis du temps à s'imposer au sein des molécularistes (après 1990) (Miklos et al., 2009). Ceux-ci peuvent de plus s'appliquer autant aux caractères morphologiques que moléculaires. La mise en application de tels modèles, en phénétique, au niveau moléculaire a entre autres permis à Woese et Fox d'approfondir les connaissances sur certaines espèces avec les critères d'homologie pris en compte par les modèles (Woese et Fox, 1977). C'est en poursuivant leurs études que Woese et ses collègues purent déterminer que certaines des espèces, les

archaebactéries, étaient suffisamment similaires entre elles et différentes des autres pour former un troisième domaine de la vie, les Archées (Woese, Kandler et Wheelis, 1990).

1.1.4 Les Archées

Les Archées, qui représentent le cœur de notre étude, sont définies comme un groupe d'organismes unicellulaires sans noyau ni organite, possédant, pour la majorité, une membrane lipidique composée de longues chaînes d'alcool isoprénoïques attachées au glycérol par des liaisons éther et contenant des types d'ARN ribosomiques qui n'existent pas chez les Bactéries ou chez les Eucaryotes (Woese, Kandler et Wheelis, 1990). Ces organismes vivent habituellement dans des conditions que l'on considérait très inhospitalières pour accueillir la vie, soit des conditions qui, pensait-on, régnaient il y a 3 à 4 milliards d'années sur notre planète, de là l'origine de leur nom. Cependant, nombre de bactéries ont été trouvées à vivre dans les mêmes conditions que les Archées. Archée provient du grec ancien « □ρχα□α », qui signifie « choses anciennes ». D'ailleurs, l'adaptation au stress énergétique chronique, provenant des conditions extrêmes diversifiées dans lesquelles elles évoluent ainsi qu'un manque de ressources énergétiques dans ces milieux, est l'un des points avec lesquels certains auteurs distinguent Bactéries des Archées (Valentine, 2007). Ce même article propose une hypothèse selon laquelle l'évolution des Archées serait dirigée par des stress environnementaux diversifiés et qui apparaissent particulièrement rapidement (Valentine, 2007).

1.2 La phylogénie des Archées

1.2.1 Particularités présentes au sein des Archées

Les Archées ont une apparence semblable à celle des bactéries. Par contre, lorsqu'on les observe plus en détail, leur contenu génétique ainsi que certains de leurs mécanismes cellulaires les rendent bien distinctes des bactéries. Entre autres, le fait que celles-ci aient certaines protéines ribosomiques ainsi qu'une membrane cellulaire différentes des Bactéries et des Eucaryotes les place dans une classe à part. Les Archées sont également très diversifiées : elles peuvent habiter dans des conditions anaérobiques et aérobiques et dans différentes conditions de température variant de 10°C à plus de

100°C. Certaines sont également aptes à vivre dans des conditions très chaudes et acides (thermoacidophiles), d'autres vivent dans des conditions extrêmement salines (halophiles) et d'autres sont aptes à la production de méthane (méthanogènes). Ces dernières sont les seuls êtres vivants capables de produire du méthane, c'est-à-dire qu'aucun Eucaryote et aucune Bactérie ne sont encore connus à posséder cette aptitude (Forterre, Brochier et Philippe, 2002). De plus, les Archées sont aptes à vivre dans bien d'autres environnements.

D'autres types de caractéristiques sont également connus pour varier chez les Archées. Elles peuvent tout aussi bien être hétérotrophes qu'autotrophes et utiliser une grande variété de donneurs et de receveurs d'électrons, par exemple $H_2+S°\rightarrow H_2S$ chez certaines *Desulfurococcales* et $4H_2+CO_2\rightarrow CH_4+2H_2O$ chez certains méthanogènes (Huber, Huber et Stetter, 2000).

Leur étude se révèle importante puisqu'avec toute leur diversité, il serait intéressant de voir dans quelles conditions et de quelle façon les différents traits les composant sont apparus et ont évolué (Forterre, Brochier et Philippe, 2002). Cette évolution des caractères phénotypiques se trouve intéressante si l'on veut également déterminer les causes génotypiques de leur apparition. Ces découvertes pourraient également être très intéressantes en ce qui a trait à l'évolution des composés biogéochimiques, taux d'oxygène et de méthane par exemple, de notre planète et l'apparition de la vie sur Terre. Certains suggèrent même que la compréhension de leurs systèmes pourrait être utile pour la découverte de la vie à l'extérieur de notre planète et pourrait avoir des implications pour les futures biotechnologies (Friend, 2007).

1.2.2 Les différents groupes d'Archées: leurs caractéristiques et leur habitat

Suite à la découverte de plusieurs de leurs membres, dont la majorité des méthanogènes, des halophiles, des thermophiles et hyperthermophiles, les Archées étaient reconnues comme étant composées de deux grands groupes : les Euryarchaeota et les Crenarchaeota. Cependant, depuis 1996, la découverte de nouvelles lignées d'Archées a amené à la création de deux autres groupes, les Korarchaeota (Barns et al., 1996) et les Thaumarchaeota (Brochier-Armanet et al., 2008), et un autre qui a

Tableau 1.1 Propriétés des Archées appartenant au phylum des Euryarchaeota. Ces données proviennent du NCBI.

	Température		Environnement		
Espèce	**Température de croissance (°C)**	**Type**	**Salinité**	**Besoins en oxygène**	**Habitat**
Aciduliprofundum boonei T469	70	Thermophile	nd	Anaérobique	Spécialisé
Archaeoglobus fulgidus DSM 4304	83	Hyperthermophile	nd	Anaérobique	Aquatique
Archaeoglobus profundus Av18 DSM 5631	> 80	Hyperthermophile	nd	Anaérobique	nd
Candidatus *Micrarchaeum acidiphilum* ARMAN-2	78	Hyperthermophile	nd	nd	nd
Ferroplasma acidarmanus fer1	40	Mésophile	nd	Anaérobique	Spécialisé
Haloarcula marismortui ATCC 43049 I	40-50	Mésophile	Halophile extrême	Aérobique	Aquatique
Halobacterium salinarum R1	50	Thermophile	Halophile extrême	Anaérobique	Spécialisé
Halobacterium sp. NRC-1	42	Mésophile	Halophile extrême	Facultatif	Spécialisé
Halogeometricum borinquense PR3 DSM 11551	40	Mésophile	Halophile	Aérobique	Marais asséché salé
Halomicrobium mukohataei DSM 12286	45	Mésophile	Halophile	Facultatif	Sol, marais salé
Haloquadratum walsbyi DSM 16790	45	Mésophile	Halophile extrême	nd	Aquatique
Halorhabdus utahensis AX-2 DSM12940	50	Mésophile	Halophile extrême	Aérobique	Terrestre
Halorubrum lacusprofundi ATCC 49239	31-37	Mésophile	Halophile extrême	Aérobique	Aquatique
Haloterrigena turkmenica DSM 5511	nd	nd	Halophile modéré	Aérobique	Spécialisé
Methanobrevibacter smithii ATCC 35061	37-40	Mésophile	nd	Anaérobique	Multiple
Methanocaldococcus jannaschii DSM 2661	85	Hyperthermophile	Halophile modéré	Anaérobique	Aquatique
Methanocella sp. RC-I	~37	Mésophile	Non-halophile	nd	Associé à un hôte
Methanococcoides burtonii DSM 6242	23,4	Mésophile	Halophile modéré	Anaérobique	Aquatique
Methanococcus aeolicus Nankai-3	42	Mésophile	nd	Anaérobique	Aquatique
Methanococcus maripaludis C5	20-45	Mésophile	nd	Anaérobique	Aquatique
Methanococcus maripaludis C6	20-45	Mésophile	nd	Anaérobique	Aquatique
Methanococcus maripaludis C7	20-45	Mésophile	nd	Anaérobique	Aquatique
Methanococcus maripaludis S2	35-40	Mésophile	nd	Anaérobique	Aquatique
Methanococcus vannielii SB	30	Mésophile	Non-halophile	Anaérobique	Aquatique
Methanococcus voltae A3	~37	Mésophile	nd	Anaérobique	Aquatique

Methanocorpusculum labreanum Z	37	Mésophile	Non-halophile	Anaérobique	Aquatique
Methanoculleus marisnigri JR1	21-25	Mésophile	nd	Anaérobique	Aquatique
Methanohalophilus mahii DSM 05219	37	Mésophile	nd	Anaérobique	Aquatique
Methanopyrus kandleri AV19	98	Hyperthermophile	Halophile modéré	Anaérobique	Spécialisé
Methanoregula boonei 6A8	37	Mésophile	nd	Anaérobique	Terrestre
Methanosaeta thermophila PT	55-60	Thermophile	nd	Anaérobique	nd
Methanosarcina acetivorans C2A	35-40	Mésophile	nd	Anaérobique	Aquatique
Methanosarcina barkeri str. Fusaro	35-40	Mésophile	nd	Anaérobique	Multiple
Methanosarcina mazei Go1	30-40	Mésophile	nd	Anaérobique	Multiple
Methanosphaera stadtmanae DSM 3091	36-40	Mésophile	nd	Anaérobique	Associé à un hôte
Methanosphaerula palustris E1-9C	30	Mésophile	nd	Anaérobique	Spécialisé
Methanospirillum hungatei JF-1	37C	Mésophile	nd	Anaérobique	Multiple
Methanothermobacter thermautotrophicus str Delta H	65-70	Thermophile	nd	Anaérobique	Spécialisé
Methanothermus fervidus DSM 2088	83	Hyperthermophile	nd	Anaérobique	Aquatique
Nanoarchaeum equitans Kin4-M	90	Hyperthermophile	nd	Anaérobique	Associé à un hôte
Natrialba magadii ATCC 43099	nr	Mésophile	Halophile extrême	Aérobique	Spécialisé
Natronomonas pharaonis DSM 2160	45	Mésophile	Halophile modéré	Aérobique	Aquatique
Palaeococcus ferrophilus DMJ DSM 13482	nd	nd	nd	nd	nd
Picrophilus torridus DSM 9790	60	Thermophile	nd	Aérobique	Spécialisé
Pyrococcus abyssi GE5	103	Hyperthermophile	nd	Anaérobique	Aquatique
Pyrococcus furiosus DSM 3638	100	Hyperthermophile	nd	Anaérobique	Aquatique
Pyrococcus horikoshii OT3	98	Hyperthermophile	nd	Anaérobique	Aquatique
Thermococcus barophilus sp nov	85	Hyperthermophile	nd	nd	Aquatique
Thermococcus kodakarensis KOD1	85	Hyperthermophile	nd	Anaérobique	Spécialisé
Thermococcus onnurineus NA1	80	Hyperthermophile	nd	Anaérobique	Terrestrial
Thermococcus sp. AM4	80	Hyperthermophile	nd	Anaérobique	Aquatique
Thermoplasma acidophilum DSM 1728	59	Thermophile	nd	Facultatif	Spécialisé
Thermoplasma volcanium GSS1	60	Thermophile	nd	Facultatif	Spécialisé

temporairement été accepté, puis inclus dans les Euryarchaeota: les Nanoarchaeota (Huber et al., 2002; Spang et al., 2010). Dans les prochains paragraphes, les aspects phénotypiques, liés notamment à la température de survie, et les habitats de chacun de ces groupes vous seront présentés. L'ordre de description de ceux-ci sera fait selon leurs traits phénotypiques, c'est-à-dire dans quelles conditions les espèces sont aptes à survivre et quelles caractéristiques prédominent chez ces espèces.

1.2.2.1 Euryarchaeota

Le nom Euryarchaeota provient du terme grec « *euryos* », qui signifie dans ce cas une « grande diversité » (Woese, Kandler et Wheelis, 1990; Brochier-Armanet et al., 2008). Au temps où ce groupe a été nommé, il n'était connu que les méthanogènes, les extrêmes halophiles, des thermoacidophiles ainsi que quelques hyperthermophiles. D'autres organismes se sont greffés à ce groupe depuis la création du phylum, tel que *Nanoarchaeum* et d'autres types de méthanogènes. La liste des génomes complètement séquencés de ces organismes, en date du 28 septembre 2009, est disponible dans le tableau 1.1.

Comme le nom du phylum l'indique, les Euryarchaeota sont très diversifiés. L'un des groupes le composant, les Halobacteriales, sont halophiles, c'est-à-dire qu'ils vivent dans des conditions extrêmement salines. Elles vivent également dans des conditions de températures mésophiles, soit entre 30 et 50 °C. Elles habitent notamment des endroits comme les étendues d'eau salée telles que la Mer Morte, les lacs salés des vallées de l'est africain et des habitats fabriqués par l'homme, comme les gisements où le sel est extrait (Kletzin, 2007). Afin de maintenir l'équilibre osmotique ainsi que de protéger les protéines contre leur dénaturation et leur précipitation due à la déshydratation, les espèces halophiles ont dû développer différentes stratégies. Les archées halophiles sont exceptionnelles dans ce cas puisqu'elles maintiennent des concentrations internes en sel isoosmotiques à l'environnement. Elles y arrivent grâce à une accumulation de KCl et de NaCl dans le cytoplasme équivalent, par exemple, à une concentration de 4.2 M de KCl et 1 M de NaCl pour l'espèce *H. salinarum*. Afin d'éviter leur dénaturation, les protéines cytoplasmiques se sont adaptées grâce à un changement dans leur composition en acides

aminés. La plupart des protéines contiennent très peu de résidus hydrophobes, mais contiennent plus de résidus acides à leur surface, ce qui augmente la solubilité des protéines dans les solutions salines. Cette capacité n'est cependant pas unique aux Archées puisque certaines bactéries possèdent la même capacité (Oren, 2002).

Un autre type de caractéristique dominante chez les Archées est l'hyperthermophilie (>80°C) ou la thermophilie (entre 50 et 80°C). Cette caractéristique est cependant également présente chez les Bactéries. Chez les Archées, les ordres possédant cette capacité sont les Archaeoglobales, les Thermoplasmatales (mis à part l'espèce *Ferroplasma acidarmanus*), les Thermococcales et les Nanoarchaeota. Il y a également des Archées méthanogènes hyperthermophiles, comme les genres *Methanopyrus*, *Methanothermus*, *Methanothermobacter* et *Methanocaldococcus*. Leurs sources d'énergie peuvent également être très diversifiées.

Afin de survivre dans de telles conditions de température, ces organismes ont dû développer différentes stratégies. Les gens croyaient que la stabilité était due à un haut niveau de nucléotide G+C (Musto et al., 2004). Cependant, ceci a été démontré comme étant plus complexe avec la disponibilité et l'analyse de génomes complets (Kreil et Ouzounis, 2001; Musto et al., 2006; Trivedi, Gehlot et Rao, 2006). Cette stabilité est due entre autres à une composition en acides aminés spécifique, qui est influencée par le taux de G+C (vrai pour les ARN stables), la propriété des reverse gyrases (Forterre et al., 2000) et à d'autres mécanismes particuliers impliquant les membranes cellulaires ainsi que la quantité d'ions et de réseaux d'ions, de ponts hydrogènes, de liaisons hydrophobes et plus d'interactions de Van der Waals (Kletzin, 2007). Les habitats typiques des organismes hyperthermophiles et thermophiles sont entre autres les geysers, les « points chauds » volcaniques tels que les cheminées marines et tout espace sous-marins à proximité de plaques tectoniques. Cette faculté n'est pas unique aux Archées cependant, car des bactéries sont aussi connues pour être aptes à vivre dans des conditions aussi extrêmes (Kletzin, 2007).

Un autre ordre, les Thermococcales sont des Euryarchaeota hyperthermophiles hétérotrophes et croissent sous des conditions anaérobiques. Elles trouvent leur énergie en effectuant de la fermentation de sucres et/ou de peptides. De plus, le soufre est requis

pour la croissance ou peut y agir comme effet stimulant. Les genres *Archaeoglobus* et *Ferroglobus* font quant à eux partie des Euryarchaeota hyperthermophiles réducteurs de sulfate ou de nitrate respectivement. Les espèces du genre *Thermoplasma*, ainsi que *Picrophilus*, sont des espèces thermoacidophiles. L'espèce *Ferroplasma acidarmanus* est acidophile, mais survit à des températures plus froides. Ces espèces résident dans des emplacements très acides comme les champs sulfureux. Ces espèces sont particulières, car elles n'ont pas de paroi cellulaire, mais seulement une membrane cellulaire qui n'est constituée que d'une seule couche lipidique composée à 80% de tétraéther. Le dernier type d'Archée hyperthermophile n'est pas encore complètement caractérisé. Il s'agit de l'espèce *Nanoarchaeum symbiosum* qui vit en symbiose/parasitisme obligatoire avec une espèce de Crenarchaeota, *Ignicoccus hospitalis*. La grandeur du génome de *Nanoarchaeum* est l'un des plus petits connus à ce jour avec moins de 600 kb (Kletzin, 2007). Une espèce d'Archée encore moins bien connue, *Candidatus Micrarchaeum acidiphilum*, qui est très petite (<500nm de diamètre), possède également des propriétés hyperthermophiles. Celle-ci est connue pour être en symbiose/parasitisme avec des espèces de *Thermoplasma* (Baker et al., 2010).

Les Archées méthanogènes sont capables de générer du méthane à partir de différentes sources, notamment du H_2 et CO_2 ainsi que du formate et de l'acétate. D'autres types de composés tels que les alcools servent également de substrats à ces espèces et vivent dans des conditions strictement anaérobiques. Les ordres Methanococcaceae, Methanosarcinales, Methanomicrobiales et quelques Methanobacteriales sont mésophiles. Une espèce, *Methanogenium frigidum* a même été trouvée dans des conditions froides, soit 15°C. Les habitats typiques pour ces espèces sont les marais d'eau douce anoxiques, les sédiments dans les océans ou les lacs, les cheminées hydrothermales, les systèmes digestifs des animaux (souvent alors endosymbiontes de protozoaires anaérobiques) (Kletzin, 2007). Aucun méthanogène n'a encore été trouvé chez les bactéries.

1.2.2.2 Crenarchaeota

Le nom Crenarchaeota provient du terme grec « *crenos* », signifiant « origine » (Woese, Kandler et Wheelis, 1990). Lorsque ce groupe a été nommé, il ne contenait que des

hyperthermophiles et son nom référait à l'origine de la vie qui, pensait-on, provenait d'organismes vivant dans ces conditions. Cependant, des organismes psychrophiles appartenant à ce groupe, c'est-à-dire vivant dans des conditions très froides, ont été découverts (Cavicchioli, 2006). De plus, différentes études ont déterminé que les Crenarchaeota étaient les organismes les plus abondants dans certains environnements marins (Wuchter et al., 2006). La liste des génomes complètement séquencés de ces organismes, en date du 28 septembre 2009, ainsi que certaines de leurs caractéristiques sont disponibles dans le tableau 1.2.

Les groupes faisant partie des Crenarcheota sont tout aussi divers que ceux faisant partie des *Euryarchaeota*. Les différents groupes sont les Thermoproteales, les Desulfurococcales, les Sulfolobales et potentiellement les Thaumarchaeota ainsi que le groupe des Korarchaeota. Les trois premiers groupes font partie sans l'ombre d'un doute des Crenarchaeota tandis que les deux derniers font encore l'objet d'un débat concernant leur position au sein des Archées. Les Thermoproteales sont des Archées hyperthermophiles et sont en grande majorité des producteurs primaires pour les environnements dans lesquels ils vivent. Ils sont d'une très grande importance dans ces environnements, qui peuvent être de nature terrestre ou marine près de sources volcaniques. Ils ont également été prélevés dans des sources d'eau acides et neutres. Certaines espèces faisant partie de ce groupe, tel que les genres *Pyrobaculum* et *Caldivirga*, sont aptes à survivre avec de faibles concentrations d'oxygène, mais la majorité d'entre elles en sont incapables. Les autres genres composant les Thermoproteales dont le génome a été séquencé sont *Thermofilum*, *Vulcanisaeta* et *Thermoproteus* (Kletzin, 2007).

Les Desulfurococcales sont des espèces également hyperthermophiles. La plupart des souches survivent dans des conditions anaérobiques, mais quelques unes sont capables de vivre dans des conditions aérobiques. Les Desulfurococcales sont séparés en deux familles, les Pyrodictiaceae, comprenant les genres *Pyrolobus* (ne fait pas partie de cette étude) et *Hyperthermus*, et les Desulfurococcaceae, comprenant les genres *Aeropyrum*, *Ignicoccus*, *Staphylothermus*, *Ignisphaera* et *Thermosphaera* (Kletzin, 2007).

Tableau 1.2 Propriétés des Archées appartenant aux phylums des Crenarchaeota, Thaumarchaeota *et* Korarchaeota. Ces données proviennent du NCBI.

Espèce	Phylum	Température		Environnement		
		Température de croissance (°C)	Type	Salinité	Besoins en oxygène	Habitat
Aeropyrum pernix K1	Crenarchaeota	90-95	Hyperthermophile	nd	Aérobique	Spécialisé
Caldivirga maquilingensis IC-167	Crenarchaeota	85	Hyperthermophile	nd	Microaérophile	Spécialisé (eau douce)
Cenarchaeum symbiosum A	**Thaumarchaeota**	10	Psychrophile	nd	nd	Associé à un hôte
Desulfurococcus kamchatkensis 1221n	Crenarchaeota	85	Hyperthermophile	nd	Anaérobique	Aquatique
Hyperthermus butylicus DSM 5456	Crenarchaeota	95-106	Hyperthermophile	Halophile modéré	Anaérobique	Aquatique
Ignicoccus hospitalis KIN4 I	Crenarchaeota	90	Hyperthermophile	nd	Anaérobique	Aquatique
Ignisphaera aggregans AQ1 S1 DSM 17230	Crenarchaeota	92	Hyperthermophile	nd	Aérobique	Aquatique
Metallosphaera sedula DSM 5348	Crenarchaeota	70	Thermophile	nd	Aérobique	Spécialisé
Pyrobaculum aerophilum str. IM2	Crenarchaeota	100	Hyperthermophile	nd	Facultatif	Aquatique
Pyrobaculum arsenaticum DSM 13514	Crenarchaeota	~95	Hyperthermophile	nd	Anaérobique	Aquatique
Pyrobaculum calidifontis JCM 11548	Crenarchaeota	90-95	Hyperthermophile	nd	Facultatif	Spécialisé
Pyrobaculum islandicum DSM 4184	Crenarchaeota	100	Thermophile	nd	Anaérobique	Spécialisé
Staphylothermus marinus F1	Crenarchaeota	92	Hyperthermophile	nd	Anaérobique	Spécialisé
Sulfolobus acidocaldarius DSM 639	Crenarchaeota	70-75	Thermophile	nd	Aérobique	Spécialisé
Sulfolobus islandicus LD85	Crenarchaeota	75-85	Hyperthermophile	nd	Aérobique	Spécialisé
Sulfolobus islandicus M164	Crenarchaeota	75-85	Hyperthermophile	Non-halophile	Aérobique	Spécialisé
Sulfolobus islandicus U328	Crenarchaeota	75-85	Hyperthermophile	nd	Aérobique	Spécialisé
Sulfolobus islandicus YG5714	Crenarchaeota	75-80	Hyperthermophile	nd	nd	nd
Sulfolobus islandicus YN1551	Crenarchaeota	75-85	Hyperthermophile	Non-halophile	nd	Spécialisé
Sulfolobus solfataricus P2	Crenarchaeota	85	Hyperthermophile	nd	Aérobique	Spécialisé
Sulfolobus tokodaii str. 7	Crenarchaeota	80	Hyperthermophile	nd	Aérobique	Spécialisé
Thermofilum pendens Hrk 5	Crenarchaeota	88	Hyperthermophile	nd	Anaérobique	Spécialisé
Thermoproteus neutrophilus V24Sta	Crenarchaeota	85	Hyperthermophile	nd	Anaérobique	Spécialisé
Vulcanisaeta distributa IC-017 DSM 14429	Crenarchaeota	85-90	Hyperthermophile	nd	Anaérobique	Spécialisé
Korarchaeum cryptofilum OPF8	**Korarchaeota**	85	Hyperthermophile	nd	Anaérobique	nd
Nitrosopumilus maritimus SCM1	**Thaumarchaeota**	28	Mésophile	nd	Aérobique	Aquatique

Les Sulfolobales sont hyperthermophiles et dépendent directement du soufre pour vivre; ils vivent alors dans des conditions acides. Ceux-ci poussent en présence d'oxygène et procèdent à l'oxydation chimolithoautotrophe du soufre (S^0), des thiosulfates, des sulfures métalliques ou le H_2. Ils peuvent également obtenir leur énergie de façon hétérotrophe avec différents substrats organiques. Les principaux genres composant les Sulfolobales sont les *Sulfolobus* et *Metallosphaeara* (Kletzin, 2007).

Les Thaumarchaeota et Korarchaeota sont deux groupes pour lesquels il existe encore des débats quant-à leur positionnement. Le nom Thaumarchaeota provient du terme grec « *thaumas* », qui veut dire « merveille » (Brochier-Armanet et al., 2008). Les trois génomes complets disponibles actuellement de ce groupe sont *Cenarchaeum symbiosum*, qui est en symbiose avec l'éponge *Axinella mexicana*, et deux espèces appartenant au type *Nitrosopumilus*, qui poussent de façon chimolithoautotrophe par nitrification. Très peu de détails sont cependant disponibles pour ces espèces (Kletzin, 2007). Le dernier groupe, faisant partie des Crenarchaeota, est les Korarchaeota, provenant du terme grec « *koros* » voulant dire « jeune homme » pour dénoter l'émergence basale de ce groupe (Barns et al., 1996). Cependant, le débat au sujet de son appartenance aux Crenarchaeota persiste, car il y a beaucoup de systèmes cellulaires présents au sein des Korarchaeota qui sont très semblables à ceux des Euryarchaeota telles que la division cellulaire, la réplication de l'ADN et la maturation des ARNt (Elkins et al., 2008), tandis que certaines de ses protéines ribosomiques sont trouvées seulement chez les Crenarchaeota, et non les Euryarchaeota (Spang et al., 2010). *Korarchaeum* possèderait peut-être la clé pour en découvrir plus sur l'ancêtre des Archées à cause de ces propriétés particulières (Elkins et al., 2008).

1.2.3 Historique de la phylogénie des Archées

Afin d'avoir une idée de la phylogénie des Archées, veuillez vous référer à la figure 1.2 où est présentée une topologie obtenue à l'aide des deux sous-unités de l'ARN ribosomique. Il existe cependant quelques mésententes par rapport au positionnement de quelques groupes, un sujet que nous traiterons au cours de ce mémoire.

Avant 1977, les Archées étaient considérées comme des bactéries parmi d'autres et ce n'est qu'avec une étude de Carl Woese (Woese et Fox, 1977) qu'une attention particulière fut lancée sur ceux-ci. La comparaison du contenu en oligonucléotides de l'ARN ribosomique 16S de différentes espèces d'Eucaryotes, de Bactéries et d'Archées est venue jeter le doute sur le paradigme de l'époque qui supposait seulement deux royaumes. Ce n'est qu'en 1990 que les Archées furent considérées comme un domaine de la vie (Woese, Kandler et Wheelis, 1990). L'ARN ribosomique fut choisi comme le marqueur de référence pour toutes les études concernant la diversité des Archées ainsi que pour les études concernant les autres domaines de la vie (Olsen et Woese, 1993). Dans cette optique où un seul marqueur est utilisé, l'augmentation de l'échantillonnage taxonomique sera la principale raison de l'amélioration des résultats obtenus (Stetter, 1996; Aravalli, She et Garrett, 1998; Buckley, Graber et Schmidt, 1998; Vetriani, Reysenbach et Dore, 1998). Le séquençage de nouvelles espèces présentes dans différents habitats et la disponibilité de plusieurs génomes complets est venu remettre en question la structure de ce groupe.

En effet, la disponibilité du génome de *Methanopyrus kandleri* et son analyse ont donné une nouvelle perspective sur l'origine de la méthanogénèse. Avec l'utilisation de l'ARN ribosomique 16S comme marqueur, *Methanopyrus* apparaissait à la base des Archées (Burggraf et al., 1991). L'utilisation de meilleures méthodes d'inférence, de méthodes différentes, tel que le contenu génomique, ainsi que l'utilisation de plusieurs marqueurs démystifia ce phénomène. *Methanopyrus* est une espèce ayant un génome riche en G+C et ce biais créait un artéfact de reconstruction en faveur d'une position basale, les protéines ribosomiques par contre ont été en mesure de placer *Methanopyrus* avec les Methanococcales et les Methanobacteriales (Slesarev et al., 2002; Brochier, Forterre et Gribaldo, 2004).

D'autres espèces, comme *Nanoarchaeum* et *Korarchaeum* ont également des caractéristiques particulières (contenu en gènes, etc.) rendant leur position phylogénétique encore incertaine. *Nanoarchaeum*, qui est la plus petite Archée connue à ce jour, a été initialement positionné à la base des Euryarchaeota (Waters et al., 2003). Sa position a par la suite été revue comme étant groupe frère des Thermococcales

(Brochier et al., 2005), une position encore hypothétique, et un nouveau groupe a même été proposé pour décrire ce genre d'organismes, les Nanoarchaeota (Huber et al., 2002).

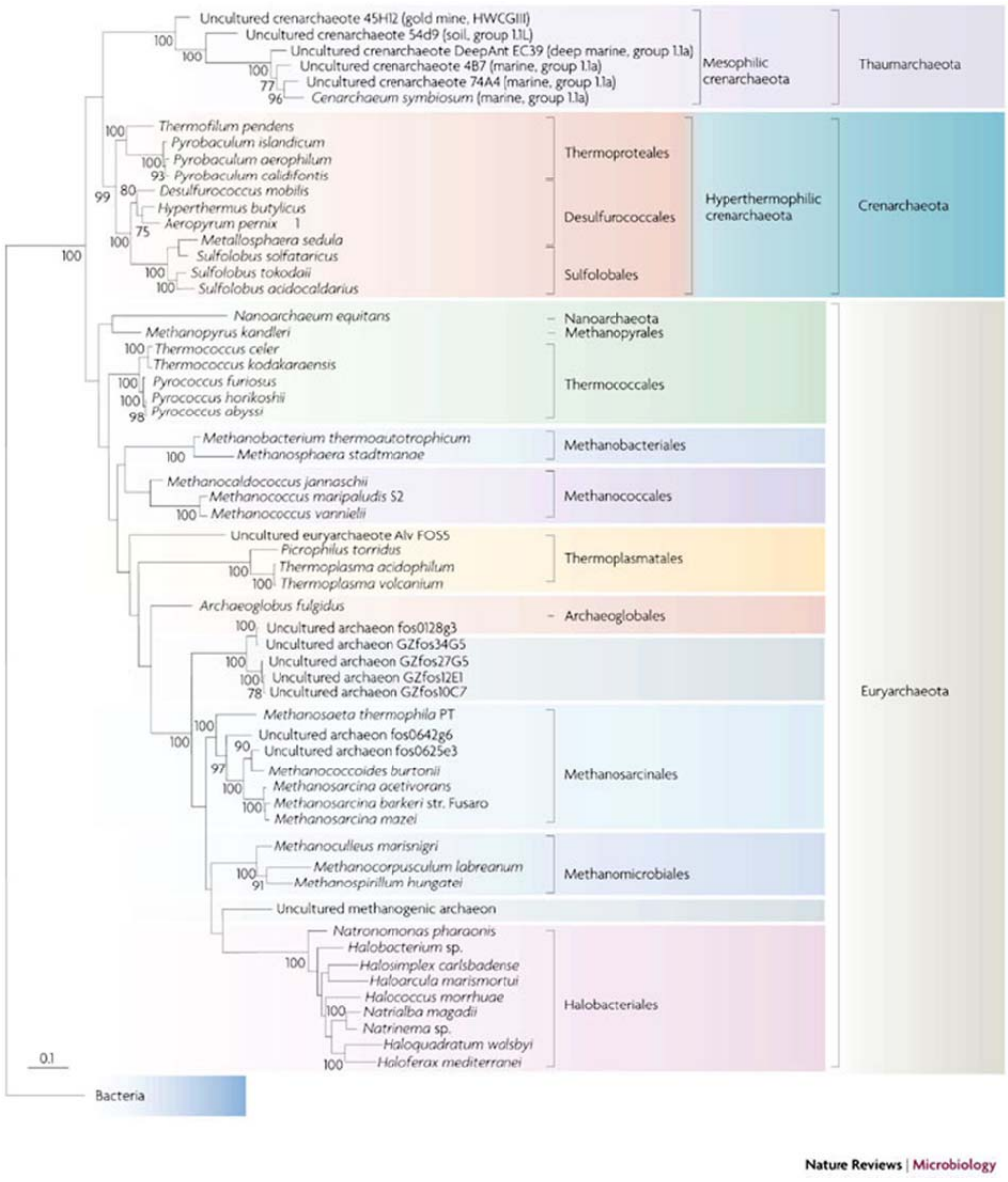


Figure 1.2 Phylogénie des Archées basée sur la concaténation des deux sous-unités de l'ARN ribosomique comprenant 3305 nucléotides et construite à l'aide de Phyml avec le modèle GTR avec une correction gamma à 8 catégories de taux de vitesse évolutive, une estimation du paramètre alpha ainsi qu'une estimation du nombre de sites invariants. Réimprimé avec la permission de Macmillan Publishers Ltd: Nature Reviews Microbiology (Brochier-Armanet et al., 2008), copyright 2008.

Les auteurs justifient qu'une position basale aux Euryarchaeota peut être retrouvée pour *Nanoarchaeum* en raison de la vitesse évolutive très rapide de cette espèce ainsi que par un biais amené par l'utilisation d'un groupe externe lointain, tel que les Eucaryotes (Brochier et al., 2005). Il se pourrait également que cette position, possiblement artéfactuelle, soit causée par un biais compositionnel. Un nouveau phylum a également été proposé pour l'espèce *Korarchaeum cryptofilum*, les Korarchaeota (Barns et al., 1996), une espèce se positionnant soit à la base des Archées, ou à la base des Crenarchaeota (Elkins et al., 2008).

Un dernier groupe vint s'ajouter en 2008, les Thaumarchaeota. Sa position est également incertaine, mais plusieurs auteurs ont obtenu, à l'aide des protéines ribosomiques ainsi que de méthodes de maximum de vraisemblance, que ceux-ci sont positionnés à la base des Archées (Brochier-Armanet et al., 2008; Spang et al., 2010). Une autre étude a proposé, avec l'aide de différents groupes externes et de méthodes plus complexes, que ceux-ci seraient le groupe frère des Crenarchaota (Foster, Cox et Embley, 2009). Il reste donc beaucoup d'incertitudes à éclaircir par rapport à la topologie de l'arbre des Archées, surtout en considérant les particularités que chacune des espèces peut avoir et les ressemblances de ces groupes par rapport aux Eucaryotes ou aux Bactéries, lorsqu'on considère par exemple le nombre de gènes partagés entre ces groupes.

1.3 Approches phylogénomiques

1.3.1 Arbre de gènes versus arbre des espèces

Il y a une grande distinction à faire en phylogénie entre la notion d'arbre d'espèces et d'arbre de gène. L'**arbre des espèces** représente l'histoire évolutive globale du génome des espèces. Une espèce, composée de plusieurs milliers de gènes, suivra une histoire évolutive spécifique et les gènes qui la composent le feront également. Cet arbre des espèces cherchera alors à trouver l'histoire de l'ensemble de gènes, reflétant l'histoire des organismes. Au contraire, un **arbre de gène** représente l'histoire évolutive des

séquences **homologues** d'un gène (qui peuvent appartenir à une ou plusieurs espèces), c'est-à-dire ayant une ancestralité commune. Des gènes homologues peuvent être de différentes natures. Il peut s'agir de gènes **orthologues**, qui sont des gènes séparés uniquement par des événements de spéciations (figure 1.3). Dans le cas où des gènes sont homologues, mais non orthologues, il peut s'agir de gènes **paralogues** ou **xénologues**. Des gènes **paralogues** sont des gènes qui ont été générés lors d'un évènement de duplication au sein d'un même organisme (figure 1.3). Des gènes **paralogues** finissent souvent par adopter des fonctions différentes ou alors l'une des copies devient inactive, celle-ci sera appelée un pseudogène. Il peut s'agir de gènes in-paralogues ou out-paralogues. Les premiers correspondent à des gènes ayant subi un ou des évènements de duplications, mais n'étant pas suivi d'évènements de spéciation, il s'agit alors de duplications du même gène survenues indépendamment sur plusieurs lignées (figure 1.3). Les seconds surviennent à la suite d'une duplication suivie de spéciations (figure 1.3). Les pertes de gènes peuvent également causer des problèmes lors de la reconnaissance des gènes orthologues et ultimement causer des problèmes lors de la reconstruction d'arbre phylogénétique des gènes ou des espèces (figure 1.4). Ces problèmes de reconstruction retrouvés majoritairement lors de la reconstruction d'arbres simple gène, résultent en des incompatibilités entre les différents arbres de gènes ainsi qu'avec l'arbre des espèces. Ces incompatibilités sont appelées **incongruences**. Des gènes homologues non orthologues peuvent également être d'une autre nature. Il peut s'agir de gènes ayant subit un transfert horizontal, ils sont alors appelés **xénologues**.

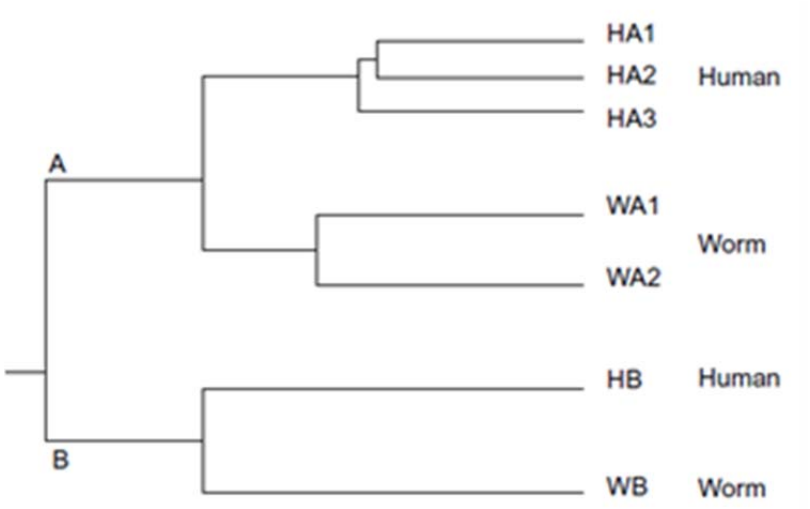


Figure 1.3 Évolution d'une famille de gènes homologues. Les gènes HB et WB sont orthologues. Les gènes HA* sont in-paralogues, tout comme les gènes WA*. Les gènes HA* et le HB par contre sont par contre out-paralogues. Adapté de Trends in Genetics, Vol 18, num. 12, (Sonnhammer et Koonin), Orthology, paralogy and proposed classification for paralog subtypes, p. 620, Copyright 2002.

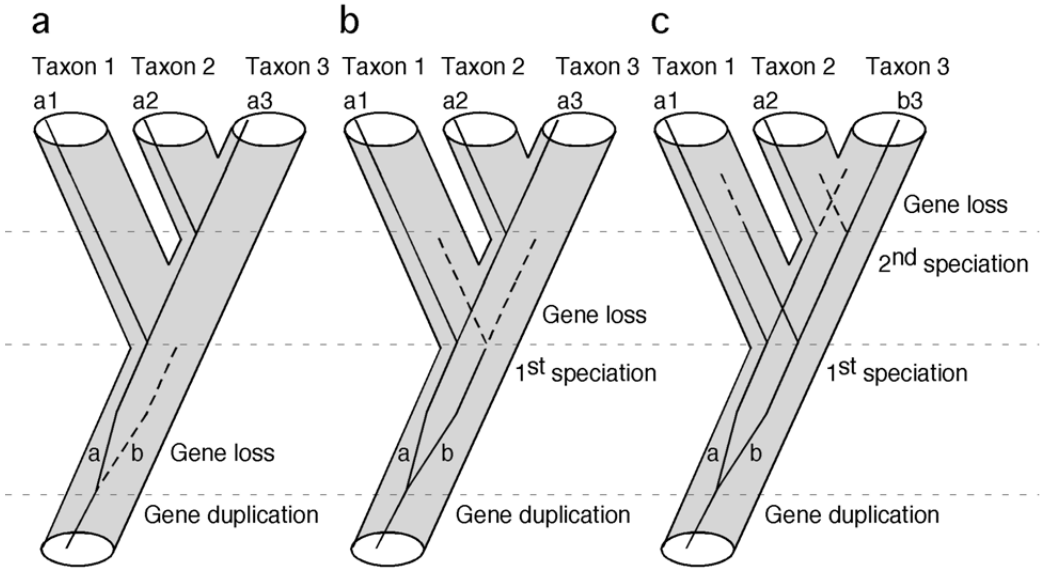


Figure 1.4 Effets des duplications suivies de plusieurs délétions de gènes sur la phylogénie obtenue (tubes recouvrant les lignes). Une duplication de gènes suivi d'une délétion avant la première spéciation (a), ou avant la seconde spéciation (b) n'affecteront pas la phylogénie résultante. L'incongruence sera créée seulement si le gène dupliqué survit les deux événements de spéciation et qu'il est perdu dans les taxas 1 et 2. Tiré de (Li et al., 2007).

Au cours de l'évolution, il peut y avoir eu chez certaines espèces, une duplication de gène, une perte ou même des transferts horizontaux de gène (Maddison, 1997). Le problème d' « incomplete lineage sorting » correspond au fait qu'un polymorphisme peut rester pendant des millions de générations dans une lignée. L'information qu'il apportera au niveau du gène dépend de la copie séquencée, alors que plusieurs copies, ayant divergées, de ce gène existent dans le génome (Maddison et Knowles, 2006). Tout cela peut mener à des problèmes lors de la reconstruction de l'**arbre des espèces** car chaque copie possèdera son propre lot de mutations différentes. Lorsqu'un grand nombre d'espèces est considéré dans une analyse simple gène et avec tous ces phénomènes existant pour chacun des gènes, la représentation de l'histoire évolutive des espèces à l'aide d'un seul marqueur n'est plus envisageable. Afin de pouvoir retrouver l'arbre des espèces, plusieurs méthodes ont été implémentées afin de passer d'arbres de gènes à arbres d'espèces. De celles-ci, il y a l'approche super-matrice et l'approche super-arbre qui seront explicitées dans la prochaine section.

1.3.2 Reconstruction phylogénomique et erreurs liées à l'homologie

Avec tous les problèmes existants au niveau des arbres de gènes, plusieurs approches ont été élaborées afin de modéliser les relations de parentés entre les différentes espèces (figure 1.5). Il existe des méthodes se basant sur des génomes complets. Certaines de ces approches ne font pas de différentiation dans la relation entre les gènes homologues (orthologues, paralogues, xénologues). Elles utilisent plutôt le signal que ces familles de gènes peuvent apporter. D'autres méthodes se focalisent quant à elles sur certaines caractéristiques, potentiellement plus fiables, des génomes, comme les gènes orthologues.

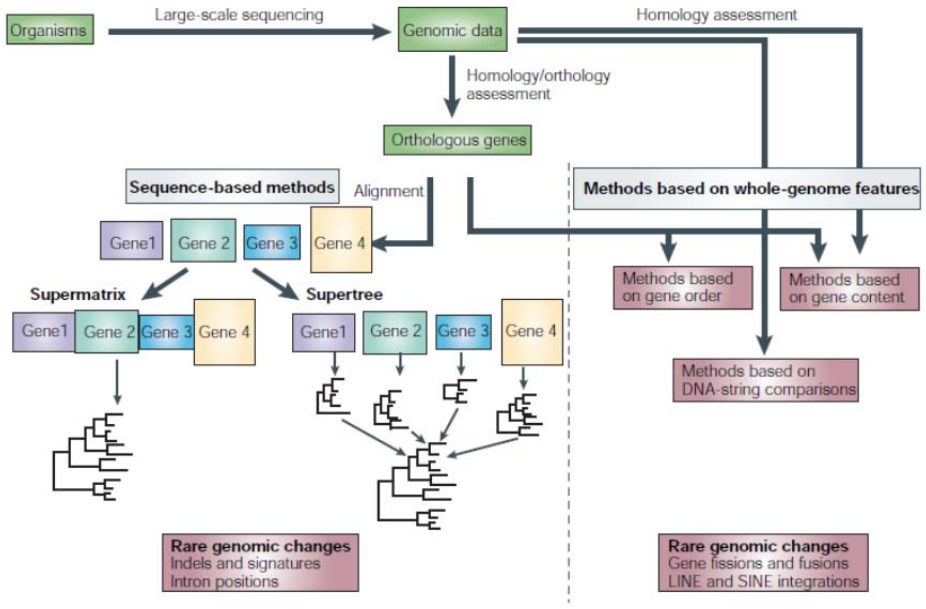


Figure 1.5 Méthodes de reconstruction phylogénomiques. Réimprimé avec la permission de Macmillan Publishers Ltd: Nature Reviews Genetics (Delsuc, Brinkmann et Philippe, 2005), copyright 2005.

Il y a des méthodes qui font des comparaisons du *contenu génomique* des différentes espèces. Ces méthodes reconstruisent les arbres phylogénétiques en se basant sur des distances qui représentent la proportion de **gènes orthologues** ou **homologues**, partagés entre les génomes (Wolf et al., 2001; Korbel et al., 2002; Delsuc, Brinkmann et Philippe, 2005). Ces distances peuvent être calculées à l'aide d'algorithmes classiques de distances (Snel, Bork et Huynen, 1999; Tekaia, Lazcano et Dujon, 1999; Lin et Gerstein, 2000) ou peuvent être calculées à l'aide de matrices où les états de caractère représentent la présence ou l'absence de gènes homologues ou orthologues dans les génomes (Fitz-Gibbon et House, 1999; Wolf et al., 2001).

D'autres approches utilisent l'*ordre des gènes* afin de reconstruire les arbres phylogénétiques. Ces approches utilisent les événements évolutifs de réarrangements de génomes, comme les duplications de gènes, les inversions, les insertions, les transpositions ainsi que les translocations. Ces méthodes visent à minimiser le nombre de changements entre les génomes (Blanchette, Bourque et Sankoff, 1997; Blanchette, Kunisawa et Sankoff, 1999). Une autre approche, qui est nommée l'approche *oligonucléotides*, n'utilise que l'information sur l'utilisation des « mots » dans le génome.

La notion d'homologie est complètement écartée pour cette approche (Lin et Gerstein, 2000; Edwards et al., 2002; Pride et al., 2003; Qi, Wang et Hao, 2004), celle-ci se base sur des combinaisons de nucléotides formant des petits mots et la fréquence de ces mots sera prise en compte, donnant ainsi de l'information sur les différentes signatures génomiques (Campbell, Mrazek et Karlin, 1999).

D'autres méthodes, basées sur les séquences, n'utilisent que des séquences orthologues. Ceci est nécessaire pour ces méthodes, qui cherchent à retrouver l'arbre des espèces, afin de n'avoir que le signal concernant les évènements de spéciation et non les duplications ou autres phénomènes. D'autres évènements peuvent causer des problèmes dans la reconstruction de l'arbre des espèces. Les gènes xénologues sont des gènes qui ont été obtenus à l'aide de transferts horizontaux, gènes provenant d'une autre espèce, qui peut être désormais éteinte. Ceux-ci causent également de l'incongruence entre les arbres de gènes et entre l'arbre des espèces et les arbres de gènes. Cependant, cette incongruence ne signifie pas qu'il s'agit d'une erreur dans la reconstruction de l'arbre de gène, il s'agit simplement que les différents arbres de gènes ont des histoires évolutives différentes. Les transferts horizontaux de gènes (THG) et leur impact sur la phylogénie seront plus amplement explicités dans la section 1.4.

Deux approches permettent d'analyser l'information provenant d'alignements de gènes orthologues. Il y a les approches super-arbres et les approches super-matrices, explicitées dans la section suivante.

1.3.3 Approches basées sur les alignements de gènes orthologues

Lorsque nous voulons effectuer une analyse en phylogénomique, nous avons le choix entre trois grandes catégories d'approches se basant sur les alignements de séquences. Cependant, ces alignements sont utilisés de différentes façons. De ces approches, la première, l'approche de bas-niveau ou **super-matrice** (figure 1.5), utilise les alignements de séquences à l'état brut. À part l'alignement, il n'y a pas de prétraitement des données. Les alignements sont concaténés, ou collés, les uns à la suite des autres afin de bâtir un seul alignement, qui sera par la suite utilisé afin d'inférer une phylogénie à l'aide d'une méthode de construction d'arbre standard. L'approche de

moyen-niveau quant-à elle calcule des informations intermédiaires telles que les distances par paires ou par quartets, et utilise ces informations pour construire un arbre (Kupczok, Schmidt et von Haeseler, 2010). L'approche de haut-niveau, ou **super-arbre** (figure 1.5), consiste à inférer tous les arbres de gène que nous désirons utiliser à l'aide d'une méthode phylogénétique standard et de les combiner ensemble afin d'obtenir un arbre des espèces (Criscuolo et al., 2006; Kupczok, Schmidt et von Haeseler, 2010). La nuance entre les approches de moyen- et de haut-niveau est que même si ces deux approches peuvent utiliser les mêmes moyens afin d'aller chercher l'information phylogénétique, ceux-ci partent de données de base différentes. L'approche de moyen-niveau utilise les alignements de séquences et va chercher directement l'information sur ceux-ci, tandis que l'approche de haut-niveau utilise des arbres phylogénétiques qui ont été inférés, avec une méthode de reconstruction phylogénétique quelconque, à partir des alignements de séquences. Le programme SDM (« super distance matrix ») peut, de par la façon dont il a été conçu, procéder de la manière moyen- ou haut-niveau par exemple (Criscuolo et al., 2006).

1.3.4 Influence du nombre de gènes et de l'échantillonnage taxonomique

Lorsque nous parlons de l'influence du nombre de gènes utilisés dans une approche phylogénomique, nous faisons indirectement référence au nombre de positions utilisées afin d'inférer l'arbre évolutif du groupe étudié. De plus, différents gènes peuvent apporter des informations différentes et le nombre de positions que ceux-ci contiennent pourra venir influer sur le résultat obtenu. En utilisant plus de gènes, l'idée est premièrement de réduire les risques d'erreurs stochastiques, liées à un trop faible nombre de positions, apportant un trop faible signal et ainsi un support statistique pour certains nœuds trop faibles. Ce type d'erreur est visible lorsque nous faisons l'étude de gènes individuels. La réduction des erreurs stochastiques ainsi que l'amélioration du support statistique des nœuds, en utilisant plusieurs gènes, furent observées par de nombreuses études empiriques différentes (Qiu et al., 1999; Soltis, Soltis et Chase, 1999; Madsen et al., 2001; Murphy et al., 2001; Bapteste et al., 2002; Rokas et al., 2003; Philippe et al., 2005).

Un autre point à ne pas négliger lorsque nous compilons des données à analyser en phylogénomique est l'échantillonnage taxonomique (Hillis, 1998; Philippe et al., 2007). En effet, celui-ci se révèle d'une très grande importance lorsque certains groupes à étudier évoluent trop rapidement par exemple, ce qui peut causer des problèmes dans la reconstruction phylogénétique, et même aussi pour avoir une meilleure estimation de l'âge des groupes lorsqu'il est question de datation moléculaire. Une sous-estimation d'un groupe combiné à une surestimation d'un autre peut engendrer des problèmes de différents ordres. Par exemple, si un groupe est représenté par une seule espèce évoluant rapidement, celle-ci aura tendance à se rapprocher d'un groupe ou d'une autre espèce rapide. Plus d'information sur un groupe monophylétique apportera, dans la plupart des cas, une meilleure estimation des temps où il y aura eu spéciation. Ceci sera possible grâce à une meilleure estimation des substitutions multiples affectant les longueurs de branches menant à ce groupe (Hendy et Penny, 1989). Il ne serait par contre pas nécessairement optimal d'avoir toutes les espèces disponibles ou même existantes. Cela pourrait dégrader l'inférence phylogénétique en incluant des espèces à évolution rapide par exemple pouvant créer des biais dans la reconstruction (Kim, 1996) ainsi que de créer de l'irrésolution si de trop nombreuse branches courtes sont présentes. De plus, dans la pratique, les résultats seraient trop longs à obtenir et les ressources nécessaires à l'obtention des résultats seraient trop importantes. Il se pourrait même qu'elles ne suffisent pas à la tâche. C'est pourquoi un choix d'espèces balancé et représentatif des groupes est si important (Stefanovic, Rice et Palmer, 2004; Philippe et al., 2005). L'utilisation d'espèces ayant une faible vitesse évolutive pour représenter les groupes à étudier accompagnés d'un groupe externe plus proche servant à raciner l'arbre peut alors être d'une grande aide (Hillis, 1998; Zwickl et Hillis, 2002; Wiens, 2005).

1.3.5 Modèles d'évolution

L'un des sujets qui sera abordé plus profondément au cours de ce mémoire sera l'utilisation de modèles d'évolution plus complexes afin d'obtenir une meilleure résolution et de régler différents problèmes liés aux modèles plus simples. Les avancées dans le domaine des modèles probabilistes évolutifs vous seront donc introduites.

Les modèles probabilistes utilisés pour les méthodes de maximum de vraisemblance pour les jeux de données nucléotidiques furent développés beaucoup plus rapidement que pour les acides aminés, car ceux-ci nécessitent beaucoup moins de paramètres (quatre états contre 20 pour les acides aminés). De nombreux modèles furent donc inventés pour l'ADN, le premier étant celui de Jukes et Cantor considérant des taux de transitions et de transversions égales pour chaque nucléotide ainsi qu'une fréquence d'équilibre égale pour toutes les bases (Jukes et Cantor, 1969). En 1980, Kimura introduisit un modèle à deux paramètres un concernant les transitions (purine (R) vers purine (R), pyrimidine (Y) vers pyrimidine (Y)) et l'autre pour les transversions (R vers Y et vice versa) (Kimura, 1980). En 1981, Felsenstein conçu un modèle pour lequel le taux de substitution correspond aux fréquences à l'équilibre des différents nucléotides (Felsenstein, 1981). Un autre modèle, le modèle HKY, combine les modèles de Felsenstein et de Kimura (Hasegawa, Kishino et Yano, 1985).

Avant 1995, les méthodes qui prédominaient pour les jeux d'acides aminés étaient principalement les méthodes de maximum de parcimonie et les méthodes de distance alors que les méthodes de maximum de vraisemblance étaient d'usage limité à cause d'un temps calcul prohibitif. Dans les années 1990, les principaux modèles utilisés pour le maximum de vraisemblance étaient assez simplistes et utilisaient des matrices de substitutions d'acides aminés, développées avec des données empiriques, représentant les propriétés chimiques physiques et biologiques de ceux-ci. Les matrices Dayhoff (Dayhoff, Eck et Park, 1972), JTT (Jones, Taylor et Thornton, 1992) et WAG (Whelan et Goldman, 2001) combinées à cette méthode considèrent une homogénéité substitutionnelle entre les sites des alignements ainsi que l'indépendance entre les sites. Une autre matrice, qui sera largement utilisée dans cette étude, LG (Le et Gascuel, 2008) est apparue dernièrement et permet une bonne amélioration dans les résultats obtenus comparativement aux autres matrices de substitution existantes car elle fut calculée à partir d'alignements très grands et une méthode de maximum de vraisemblance incorporant la variabilité des taux évolutifs le long des sites. Un modèle considérant l'homogénéité substitutionnelle entre les sites insinue que tous les sites d'un alignement évoluent selon le même patron de substitution.

Le problème de correction des différents taux d'évolution entre les sites devenait alors l'un des problèmes prioritaires à régler. La distribution gamma a alors fait son entrée afin de résoudre une partie des problèmes générés par cette propriété (Nei et Gojobori, 1986; Tamura et Nei, 1993; Yang, 1993), mais a mis tout de même du temps à se mettre en place dans la communauté notamment à cause de problèmes d'implémentation. Yang a par la suite démontré qu'un modèle gamma discret obtenu avec quatre catégories de taux évolutifs était un bon compromis entre le temps calcul et la précision des résultats obtenus pour être utilisé comme une approximation d'une distribution gamma (Yang, 1994; Lio et Goldman, 1998). Ces méthodes permettaient d'obtenir des estimations des phylogénies simple gène et multi-gènes relativement rapidement. Beaucoup de problèmes dans la reconstruction et d'erreurs étaient cependant obtenus avec celles-ci car les modèles simples ne sont pas aptes à décrire correctement les jeux de données.

1.3.6 Erreurs stochastiques et systématiques

La reconstruction phylogénétique n'est pas à l'abri de différents types d'erreur. Il y a les erreurs stochastiques, lorsqu'il n'y a pas assez de sites dans l'alignement, ce qui fait en sorte que le signal est trop faible pour reconstruire l'arbre avec suffisamment de résolution. Certaines erreurs surviennent dans la reconstruction lorsqu'une inférence atteint un minimum local par exemple. Il s'agit dans ce cas d'une erreur d'heuristique dans la reconstruction. Ce genre d'erreur survient souvent alors que la position d'une espèce est séparée de sa bonne position par plusieurs nœuds. Ce type d'erreur est en particulier dû à une méthode de recherche de l'espace des arbres appelé NNI (Waterman et Smith, 1978), pour « nearest neighbor interchange ». Il s'agit d'une méthode de recherche de topologie qui consiste à interchanger deux voisins proches, donc à faire des réarrangements locaux, jusqu'à atteindre un minimum dans le logarithme de la vraisemblance du jeu et cela selon le modèle d'évolution utilisé (Page, 1993). Pour cette raison, la plupart des méthodes utilisent désormais des réarrangements topologiques affectant des branches sur l'arbre en entier plutôt que des réarrangements sur des parties locales de l'arbre.

Dans un cadre strictement probabiliste, il y a d'autres erreurs qui sont liées à des violations du modèle utilisé, les erreurs systématiques. Ces erreurs proviennent du fait que la méthode de reconstruction ou le modèle d'évolution utilisé ne reflète pas bien le processus utilisé du jeu de données analysé. Ces erreurs sont de plus en plus soutenues avec l'augmentation du nombre de sites utilisés. Ces erreurs se manifestent par exemple comme une attraction des longues branches (voir la section 1.3.8) pour les méthodes de maximum de parcimonie. Pour les méthodes probabilistes, les erreurs systématiques sont exclusivement dues à des violations de modèles (Douzery et al., 2010). Afin de réduire ces erreurs de violations du modèle, il n'y a d'autres solutions que d'améliorer ces modèles afin qu'ils puissent prendre en compte le plus de caractéristiques évolutives que possible contenues dans les jeux de données.

1.3.7 Améliorations apportées par les modèles d'évolution de séquence

Avec l'approche super-matrice, l'utilisation de davantage de positions contenant de l'information phylogénétique ainsi que l'amélioration de la puissance de calcul, de par le nombre de processeurs utilisés ainsi que de leur rapidité grandissante, a permis l'utilisation de nouveaux modèles d'évolution plus complexes, utilisés tant pour les méthodes de maximum de vraisemblance ou d'inférences bayésiennes. La matrice GTR, pour « General Time Reversible », a apporté déjà quelques amélioration pour les modèles standards en permettant d'aller chercher plus d'information sur les caractéristiques substitutionnelles et biologiques du jeu de données étudié (Lanave et al., 1984). La matrice GTR peut effectivement être calculée sur les jeux nucléotidiques ou protéiques, cependant, plus la matrice à déterminer est grande, plus nous aurons de paramètres à estimer. Les modèles complexes, différents des modèles standards, prennent désormais en compte plus de caractéristiques évolutives et ajoutent de nombreux paramètres à inférer dans les analyses, pouvant aider à réduire le signal non phylogénétique pour la construction d'arbres phylogénétiques. Ils permettent d'aller chercher du signal évolutif, ignoré avec les modèles plus simples, et d'approfondir la recherche sur les données disponibles à l'aide de plusieurs paramètres supplémentaires, représentant des caractères évolutifs différents.

Entre autres, le modèle CAT (Lartillot et Philippe, 2004) est un modèle de mélange qui considère l'hétérogénéité substitutionnelle entre les sites. Ce modèle, pour lequel les probabilités stationnaires, stables au cours du temps, sont site-spécifique, permet de définir différents profils substitutionnels des sites de l'alignement représentant différentes positions contigües de l'alignement (Le, Lartillot et Gascuel, 2008). Afin d'alléger le calcul, différentes stratégies existent et celle empruntée par le modèle CAT consiste à regrouper les sites ayant des patrons substitutionnels similaires, réduisant de cette façon drastiquement le nombre de paramètres à estimer. Cette méthode permet d'aller chercher plus d'information sur l'histoire mutationnelle des séquences et elle est capable de détecter plus facilement des substitutions multiples, non possible avec des modèles considérant l'homogénéité entre les sites (Lartillot, Brinkmann et Philippe, 2007; Baurain et Philippe, 2010). Ceci permet de minimiser l'incidence de certains artéfacts de reconstruction des arbres tels que l'attraction des longues branches (Brinkmann et al., 2005; Lartillot, Brinkmann et Philippe, 2007). Les substitutions multiples sont alors beaucoup mieux cernées avec les modèles hétérogènes que les modèles considérant l'homogénéité du taux de substitution entre les sites. D'autres modèles ont également été implémentés considérant l'hétérogénéité substitutionnelle des différents sites d'un alignement. La vitesse évolutive des différents sites d'une protéine ne sont pas indépendante et celle-ci dépend entre autres de facteurs tels l'accessibilité au solvant, le code génétique, les structures secondaire et tertiaire, la fonction de la protéine, etc. La meilleure façon d'obtenir de l'information évolutive sur les sites d'une protéine serait d'établir un modèle par position. Par contre, ceci n'est pas statistiquement possible, il y aurait trop de paramètres à estimer. Lorsque la catégorie d'un site en particulier concernant des caractéristiques précise est cependant connue (ex. l'accessibilité au solvant), des méthodes de partitionnement (différentes des modèles de mélanges) analyseront chaque position avec la matrice de substitution d'acides aminés appropriée pour sa caractéristique (Felsenstein et Churchill, 1996; Thorne, Goldman et Jones, 1996; Le, Lartillot et Gascuel, 2008). D'autres modèles se concentrant plus particulièrement sur les codons ont également été implémentés. Ceux-ci modélisent tout ce qui a trait aux substitutions, soient synonymes ou non-synonymes, et les composants sélectifs des processus de substitution et modélisent également, pour certains modèles,

l'interdépendance des sites (Tillier et Collins, 1995; Rodrigue, Philippe et Lartillot, 2006; Delport, Scheffler et Seoighe, 2009).

D'autres modèles cherchent à résoudre le problème d'hétérotachie, qui décrit le fait que les vitesses d'évolution varient non seulement à travers les sites, mais également à travers le temps (Lopez, Casane et Philippe, 2002). Afin de résoudre les problèmes que pouvaient engendrer un tel processus évolutif, le modèle d'évolution de séquence covarion a été proposé (Fitch et Markowitz, 1970). L'hypothèse covarion stipule qu'à un temps précis au cours de l'évolution, étant donné certaines contraintes fonctionnelles, différents sites d'une protéine peuvent varier et d'autres non. Cette hypothèse recrée donc le principe d'hétérotachie de façon indirecte et plus simplement. Plusieurs modèles, basés sur l'hypothèse covarion ont été implémentés (Tuffley et Steel, 1998; Galtier, 2001; Huelsenbeck, 2002; Wang et al., 2007). Ceux-ci se basent sur différents principes que je n'évoquerai pas dans ce mémoire.

Dans ce mémoire, nous utiliserons principalement le modèle CAT ainsi que les matrices GTR, qui sont plus flexibles et qui permettent d'aller chercher plus d'informations évolutives dans nos jeux de données qui sont assez différents de jeux d'eucaryotes (voir chapitre 2).

1.3.8 Artéfacts de reconstructions

Il existe de nombreuses manifestations d'erreurs de reconstruction phylogénétique. Tel que mentionné dans les sections 1.3.6 et 1.3.7, l'attraction des longues branches (ALB) est l'une de ces manifestations. Cette erreur survient lorsque deux taxons non apparentés évoluent rapidement. Ce taux élevé de changements augmente la probabilité d'homoplasies, qui sont des états de caractères identiques non issus d'un ancêtre commun, et le résultat de cela est que si ces deux taxons présentent des branches terminales beaucoup plus longues que leurs plus proches parents, ceux-ci seront attirés l'un vers l'autre. Sous certaines conditions, ce résultat est le plus parcimonieux (Felsenstein, 1978). Cet artéfact est d'autant plus problématique lorsque les longues branches sont à la base de l'arbre en présence d'un groupe externe lointain.

Les espèces évoluant rapidement sont alors attirées vers la racine de l'arbre (Brinkmann et al., 2005).

Les biais compositionnels peuvent également créer des artéfacts de reconstruction. Les séquences d'espèces éloignées ayant un taux de G+C similaire seront groupées ensemble dans l'arbre inféré par une méthode ne prenant pas en compte ce problème (Woese et al., 1991; Lockhart et al., 1994).

1.3.9 Signal phylogénétique et non-phylogénétique

Au cours de l'évolution, les gènes subissent des substitutions le long de leur séquence et c'est ce qui représente le signal phylogénétique qui sera utilisé pour la reconstruction des arbres. Cependant, l'accumulation de ces substitutions à une même position peut créer des problèmes lors de la reconstruction phylogénétique, surtout pour la reconstruction des nœuds et branches les plus profonds. Lorsque le nombre de substitutions multiples excède le nombre de substitutions simples pour un certain marqueur, on dit que le jeu de données est saturé pour les espèces considérées (Baurain et Philippe, 2010). Si les branches internes sont déjà courtes, les substitutions multiples donneront généralement des arbres non-résolus. Cette conséquence n'est cependant pas la plus grave. Lorsque ces substitutions donnent de l'homoplasie, et que celles-ci sont plus abondantes que les caractères dérivés partagés, les synapomorphies, les méthodes phylogénétiques peuvent converger vers des solutions fausses hautement supportées. Les méthodes probabilistes ont en partie été conçues afin de remédier à cette situation et permettre de capter et d'inférer ce signal ancestral (Baurain et Philippe, 2010). Dans un monde idéal, ces méthodes devraient être protégées de ces problèmes de saturation puisqu'elles sont conçues pour imiter les processus évolutifs. Ceux-ci sont par contre si complexes et demanderaient tellement de ressources informatiques qu'ils sont énormément simplifiés. Certaines stratégies pour contourner ces problèmes ont cependant été élaborées. Le modèle CAT par exemple (voir section 1.3.4), a été conçu pour éliminer une grande part du signal non-phylogénétique causé par les substitutions multiples, entre autres (Lartillot et Philippe, 2004; Lartillot, Brinkmann et Philippe, 2007; Baurain et Philippe, 2010).

1.4 Historique et biologie des transferts horizontaux de gènes (THG)

Les problèmes liés à l'utilisation des modèles d'évolution de séquence sont une chose et seront plus approfondis lors du chapitre 2 de ce mémoire. D'autres types de problèmes peuvent également survenir lors de la reconstruction de l'arbre des espèces. Chez les procaryotes plus particulièrement, il existe le problème des transferts horizontaux de gènes (THG). Ce deuxième type de problème de reconstruction sera l'élément principal discuté lors du chapitre 3.

1.4.1 Découverte du THG

Le phénomène de transfert de gène fut remarqué lorsque des chercheurs constatèrent que certains phénotypes se répandaient parmi les cultures bactériennes, suggérant alors un échange de matériel génétique entre individus, ne survenant pas seulement lors des cycles de réplication. Le premier cas répertorié fut celui de la transmission de la virulence chez les pneumocoques chez des souris infectées (Griffith, 1928). Les transferts horizontaux de gènes furent également caractérisés en 1959 par un groupe japonais qui découvrit au hasard ce phénomène (Davies et Davies, 2010). Ceux-ci ont en effet constaté qu'une résistance aux antibiotiques était acquise si facilement par différentes souches d'une même espèce de bactéries qu'il était impossible que celles-ci génèrent *de novo* la résistance et que l'explication la plus raisonnable était un transfert de gènes entre les lignées, ou souches (Ochman, Lawrence et Groisman, 2000). Cette idée fut accueillie avec beaucoup de scepticisme en occident, mais changea profondément le paradigme en microbiologie quant à la résistance aux antibiotiques (Davies et Davies, 2010). L'impact de ceux-ci sur l'évolution ne fut apprécié que beaucoup plus tard et plusieurs courants de pensée ont suscité et suscitent encore beaucoup de débats. L'impact des transferts est certes assez important en ce qui concerne les procaryotes (Jain, Rivera et Lake, 1999), et certains indices laissent également croire que les THG ont une certaine importance dans les processus évolutifs chez les protistes (Bapteste et al., 2005). Cependant, le concept était plus difficile à imaginer pour les eucaryotes multicellulaires. Plusieurs exemples ont été trouvés chez les plantes et animaux (Mae-Wan Ho, 1999), mais l'importance de tels mécanismes chez ces organismes reste encore

très incomprise et la rareté de ces évènements chez les eucaryotes multicellulaires contribue à cette incompréhension (Richardson et Palmer, 2007).

1.4.2 Mécanismes du THG

Il existe trois principaux mécanismes biologiques pour la transmission horizontale des gènes: la transformation, la transduction et la conjugaison. La transformation implique une saisie de matériel génétique libre dans l'environnement par son nouvel hôte et a comme possibilité de pouvoir transmettre ce matériel entre deux hôtes très éloignés évolutivement. La transmission de matériel génétique peut également s'effectuer à l'aide d'un bactériophage qui aurait emmagasiné au hasard certains fragments d'ADN de son premier hôte, ce qu'on appelle la transduction. La transmission de ce matériel vers le second hôte du bactériophage ne nécessite pas que le donneur et receveur soient proches dans l'espace ou même dans le temps. Les protéines du phage offrent une bonne protection pour les séquences qui seront transférées et permettent, en plus du transfert dans le cytoplasme du receveur, la possible intégration chromosomale. Le troisième mécanisme, la conjugaison, implique la proximité physique du donneur et du receveur, puisque le transfert se fera via un plasmide mobilisable (Ochman, Lawrence et Groisman, 2000). Pour que le transfert s'effectue, il y a formation d'un pilus, qui est une jonction entre les deux cellules. Deux autres étapes sont nécessaires pour la conjugaison. La seconde étape consiste à effectuer de la signalisation afin de faire débuter le transfert et finalement le transfert (Frost et al., 2005). La conjugaison permet entre autres de pouvoir effectuer des transferts entre des procaryotes et des Eucaryotes multicellulaires, comme les plantes (Buchanan-Wollaston, Passiatore et Cannon, 1987), ou avec la levure par exemple (Heinemann et Sprague, 1989).

Après l'évènement de transfert, le gène ou les gènes étrangers doivent être intégrés dans le génome de l'organisme hôte. Si le matériel génétique étranger possède une similarité assez élevée avec une partie du génome hôte, celui-ci peut s'intégrer par recombinaison homologue et alors remplacer la séquence originale et sinon, par recombinaison illégitime auquel cas le génome sera devenu plus grand. La transmission par plasmide représente la seule exception, puisque les plasmides sont capables de

s'auto-répliquer en utilisant la machinerie de la cellule hôte grâce à une origine de réplication qui leur est propre (Thomas et Nielsen, 2005).

1.4.3 Fixation dans la population

Les chances qu'un gène transféré soit fixé dans une population composée de millions d'individus, c'est-à-dire qu'il se répande chez tous les individus, sont assez faibles, il faudrait même des millions d'années pour certains gènes à être fixés (Lercher et Pal, 2008). Un gène doit être premièrement transféré dans une seule cellule pour par la suite se fixer grâce à l'avantage sélectif qu'il pourrait conférer ou encore grâce à la dérive génétique.

Cependant, plusieurs facteurs s'opposent à la fixation d'un ou plusieurs gènes dans une population. Les gènes qui n'apportent pas ou très peu d'avantage sélectif auront une faible probabilité d'être fixés, cette probabilité sera inversement proportionnelle à la taille de la population, qui est énorme chez les procaryotes (Kimura, 1962). La plupart du temps, les séquences transférées seront neutres ou quasi-neutres (Berg et Kurland, 2002; Keeling et Palmer, 2008). Les mutations aléatoires ainsi que le coût de maintien de ces nouveaux gènes sont également des facteurs s'opposant à la fixation des gènes qui pourraient entrainer une perte de compétitivité des individus au sein d'une population. Une adaptation des organismes pourrait être alors de mettre moins d'énergie pour le maintien de son génome, et donc de se débarrasser de ces gènes surnuméraires. Ces mécanismes de purification font partie d'un processus évolutif, appelé nettoyage génétique (Berg et Kurland, 2002). De plus, la fixation d'un gène homologue dans une population est très improbable puisque ce gène aura été conçu préalablement pour être efficace dans son hôte original, ce qui ne signifie pas qu'il le sera plus que le gène indigène de la nouvelle population (Berg et Kurland, 2002). S'il advient que ceux-ci se fixent, ils auront plus de chance d'obtenir une nouvelle fonction par mutation que de conserver la même fonction et de remplacer la protéine existante, si elle existe, dans l'organisme hôte (Kurland, Canback et Berg, 2003).

Selon Hao *et al.* (2006), un grand nombre de gènes transférés sont présents chez les espèces actuelles, donc à l'extrémité de l'arbre. Ces gènes évolueraient plus

rapidement et d'une façon plus flexible que les gènes indigènes. Il s'agirait d'évolution directionnelle qui favorise l'intégration et l'harmonisation du gène dans le génome. Ces gènes seraient alors selon eux sélectionnés rapidement afin de fournir un avantage sélectif pour notamment favoriser l'adaptation dans une nouvelle niche écologique. Ces gènes seraient perdus rapidement au cas où l'espèce changerait de niche écologique. Il serait alors plus question dans ce cas-ci de gènes passagers plutôt que de gènes s'établissant pour de bon chez l'espèce (Hao et Golding, 2006). Dans le cas où le gène n'apporte pas d'avantage sélectif, le gène en question mutera également rapidement, parce qu'ils ne seront pas nécessaires à leur nouvel hôte et pourraient être en voie de disparaitre du génome. Le nombre de facteurs s'opposant à la fixation d'un gène transféré ainsi que ces phénomènes d'adaptation rapide à des niches écologiques laissent présumer que la fixation de transferts pourrait être plus difficile que certains le prétendent. Ce modèle qu'ils ont élaboré a également démontré que la majorité des gènes transférés disparaissent très vite à cause de tous ces facteurs d'opposition.

1.5 Fréquence des THG et remise en question de l'arbre de la vie

1.5.1 Fréquence et type de gènes affecté

Plusieurs études se sont consacrées à l'évaluation de la fréquence des THG au cours de l'évolution. Il est difficile de différencier ceux-ci de la perte de gènes (Ragan, 2001; Lawrence et Ochman, 2002; Ragan, 2002). Bon nombre de ces études suggèrent que les THG ont eu un impact important au cours de l'évolution, mais d'autres contredisent ces résultats, tout dépendant si le but de la méthode était de considérer les pertes de gènes ou les transferts (Faguy et Doolittle, 1999; Nelson et al., 1999; Ochman, Lawrence et Groisman, 2000; Ochman, 2001; Kurland, Canback et Berg, 2003; Philippe et Douady, 2003; Koonin et Wolf, 2008). La conclusion est que les THG jouent indubitablement un rôle important dans l'évolution des espèces, mais la fréquence de ceux-ci est de loin inférieure aux transmissions verticales, ou de cellules mères aux cellules filles, des gènes (Kurland, Canback et Berg, 2003). À l'aide de critères de similarité entre les gènes, il est estimé que 1,6% à 32,6% des gènes de chaque génome microbien a été acquis par THG (Koonin, Makarova et Aravind, 2001). L'impact cumulatif de ceux-ci à travers l'arbre de l'évolution mènerait à un taux de transfert

encore plus élevé, soit un impact dramatique de 81±15% (Dagan, Artzy-Randrup et Martin, 2008). D'autres études suggèrent cependant le contraire. L'une d'entre elles utilise de l'information sur les familles de protéines et utilise une méthode de comparaison d'arbres phylogénétiques afin de détecter les THG et vient à la conclusion que le taux de transferts est plutôt faible, soit 0% à 22%, et que l'impact des THG sera très faible dans la reconstruction phylogénétique (Choi et Kim, 2007). Des résultats aussi contradictoires concernant la fréquence des THG ainsi que leur impact sur les arbres d'espèces inférés démontrent l'ampleur du débat encore existant par rapport aux THG.

Avec la différence de complexité des différents systèmes biologiques, comme les voies métaboliques par exemple, certains gènes ne devraient pas être aussi susceptibles que d'autres à être transférés. L'hypothèse de la complexité (Jain, Rivera et Lake, 1999) émet que plus un gène est impliqué dans un grand nombre d'interactions dans la cellule, plus celui-ci sera difficile à remplacer puisqu'il est déjà très bien adapté à son hôte. De plus, celui-ci sera important pour la survie de la cellule et il serait moins probable qu'il intervienne dans un évènement de transfert sélectionné ou viable. Cette hypothèse fut vérifiée et les taux de transferts selon les catégories des gènes furent quantifiés. Par exemple, les gènes impliqués dans la traduction, le trafic intracellulaire ainsi que la transcription sont moins susceptibles que des gènes impliqués dans le métabolisme à être transférés (Cohen et Pupko, 2010; Puigbo, Wolf et Koonin, 2010).

1.5.2 L'impact de la proximité des espèces et le type d'habitat

Certains auteurs proposent que la proximité évolutive des espèces rende favorable les THG entre elles, et créerait un biais de transferts entre différentes espèces (Gogarten, Doolittle et Lawrence, 2002). La proximité évolutive de ces espèces impliquerait que ces espèces soient aptes à survivre dans les mêmes milieux. Les transferts entre espèces lointaines pourraient également survenir grâce à leur présence dans un même environnement. Par exemple, des transferts entre Bactéries et Archées se produiraient en grande partie à cause de leur proximité physique, parce qu'ils vivent dans le même habitat (Beiko, Doolittle et Charlebois, 2008).

1.5.3 Remise en question du concept d'espèce chez les bactéries et les archées

Chez les procaryotes en général, où le taux de THG pourrait être potentiellement élevé, certains auteurs ont redéfini la notion même d'espèce. La notion d'espèce chez les Procaryotes et Eucaryotes unicellulaires serait différente de celle des Eucaryotes multicellulaires puisque les frontières pour délimiter une espèce est plus floue chez les premiers (Gogarten et Townsend, 2005). La notion d'espèce est ambigüe à cause entre autres du nombre d'événements de THG possibles chez ceux-ci et à cause que la transmission et l'intégration des gènes par recombinaison peut se faire entre différentes espèces (Gogarten et Townsend, 2005). Ces auteurs affirment notamment que la notion d'espèce devrait s'appliquer à un niveau taxonomique plus élevé chez les Procaryotes puisqu'un grand nombre de Procaryotes s'échangent préférentiellement des gènes et ce même s'ils sont à des distances évolutives relativement très grandes. La signification même de l'arbre binaire et de ce qu'il indique se trouve affecté par cette redéfinition dans le cas où des THG préférentiels s'effectuent entre différentes espèces. Les THGs, plus fréquents que les transferts verticaux selon ces auteurs, pourraient dans ce cas devenir le signal phylogénétique et remplacer la notion de parenté au sens strict de l'ancestralité (Gogarten, Doolittle et Lawrence, 2002).

1.5.4 Remise en question de l'utilisation d'un arbre

Telles que mentionnées dans la section 1.5.2, les préférences pour que des THG soient survenus entre espèces proches ont poussé certains, comme WF Doolittle (1999), à remettre le concept d'arbre en question, du moins pour les procaryotes. Celui-ci émet l'hypothèse que l'histoire évolutive de la vie devrait être représentée comme un réseau plutôt qu'un arbre. Plusieurs autres modèles, plus flexibles qu'un arbre, sont également proposés. Un de ces modèles, le « cobweb » de la vie, propose de prendre l'arbre de l'évolution, étant représenté par des liens plus épais, et d'y ajouter des liens représentant les THG, en connexions plus minces (Ge, Wang et Kim, 2005). Bapteste *et al.* proposent quant à eux le modèle de synthèse de la vie qui considère également l'arbre d'évolution classique, mais dont certains nœuds seraient reliés de façon horizontale, référant aux grandes routes de THG (Bapteste et al., 2004). Olendzenski et Gogarten proposent que

l'évolution devrait plus être vue comme un « corail » ou un germe de pomme de terre, avec énormément de transferts près de la racine, et moins d'évènements plus nous avançons vers les branches terminales (Olendzenski et Gogarten, 2009). Tous ces modèles peuvent être construits à l'aide de réseaux. Cependant, ceux-ci étant trop lourds en temps calcul, difficilement lisibles et peu développés pour des applications en phylogénie, les phylogénéticiens utilisent encore majoritairement les arbres.

1.5.5 Effet des THG sur la phylogénie simple gène

L'incongruence obtenue chez les procaryotes peut être due à plusieurs causes. Il peut s'agir de duplications de gènes suivies ou non de pertes de gènes (figure 1.4) ou encore de THG. Sur la figure 1.6, nous voyons ce que cause un transfert récent sur une phylogénie simple gène. Le transfert aura comme conséquence sur une phylogénie d'associer l'espèce receveuse avec l'espèce donneuse, comme s'il s'agissait de deux espèces fortement apparentées. Les transferts plus anciens, c'est-à-dire ceux qui surviennent plus près de la racine, sont quant à eux beaucoup plus difficiles à identifier, voire impossibles, si trop de transferts ou d'évènements de pertes et de duplications de gènes se sont produits au cours de l'évolution pour une famille de gènes particulière.

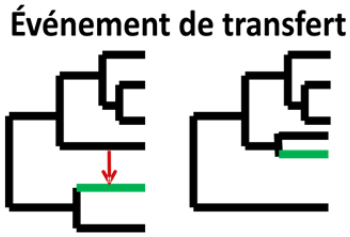


Figure 1.6 Conséquences d'un transfert horizontal sur la phylogénie simple gène obtenue. L'arbre de gauche représente l'arbre des espèces sur lequel un évènement de transfert survient entre deux espèces contemporaines et l'arbre de droite l'arbre simple gène résultant.

1.5.6 Méthodes de détection des THG

Avec le genre de conséquences qu'un transfert de gène peut avoir sur une phylogénie simple gène, certaines techniques ont été créées afin de pouvoir déterminer les séquences de gènes homologues qui sont affectés par ceux-ci. Il est possible de ce

fait de calculer la fréquence des THG parmi différentes catégories de gènes. Afin d'en être capables, différentes techniques de détection de THG ont été élaborées. Les plus traditionnelles sont appelées méthodes phylogénétiques de détection et il y a également les méthodes non-phylogénétiques. Les méthodes phylogénétiques nécessitent, comme leur nom l'indique, la construction d'arbres phylogénétiques pour chaque gène et l'incompatibilité (ou discordance, incongruence) est détectée à l'aide d'un arbre des espèces de référence. L'arbre de référence est généralement construit à l'aide de gènes universels, comme les protéines ribosomiques ou la SSU ARNr (Brochier et al., 2002; Daubin, Lerat et Perriere, 2003). Les méthodes non-phylogénétiques comptent de leur côté sur les différentes caractéristiques des génomes des organismes afin de pouvoir détecter des régions anormales présentes ayant été potentiellement « victimes » de transferts (Poptsova, 2009).

1.5.7 Méthodes phylogénétiques

Les méthodes de détection phylogénétiques de THG sont appelées traditionnelles, car il s'agit de la façon la plus instinctive pour détecter des anormalités parmi les gènes. Il s'agit de comparer les arbres de gènes homologues avec l'arbre des espèces et ainsi, mettre en évidence des incongruences, contredisant l'histoire verticale de l'arbre de référence. Certaines de ces incongruences peuvent être dues à autre chose qu'un THG, par exemple un gène paralogue, une perte de gènes ou des erreurs de reconstruction (stochastiques ou systématiques). Lorsque ces raisons d'incongruences peuvent être exclues, nous sommes alors en présence d'un THG.

Ce qui est observé dans ce cas, c'est que le gène xénologue (ayant subi un THG), sera plus proche parent du gène de l'espèce donneuse que des gènes orthologues de ses plus proches parents selon l'histoire des espèces. Plusieurs méthodes ont été élaborées afin de détecter les THG ainsi que d'établir la direction de ceux-ci. Nombre de celles-ci utilisent le concept de réconciliation d'arbres qui a été introduit par Mirkin *et al.* en 1995 (Mirkin, Muchnik et Smith, 1995). Cette méthode consiste à comparer toutes les topologies simple gène et cherche à les combiner en un seul arbre des espèces par un minimum d'évènements de changements topologiques (par différents types de mouvements de branches) ou de retrait de taxons. Ce problème a été interprété de

plusieurs façons, certains cherchent le MAF (« maximum agreement forest »), problème qui a été imaginé en 2007 (Rodrigues, Sagot et Wakabayashi, 2007). Ceux-ci cherchent à trouver le nombre minimal de branches à couper dans chaque arbre afin d'obtenir deux forêts de sous-arbres racinés. Une autre interprétation du problème, le SPR (« subtree pruning and regrafting »), va également dans le même sens que MAF (Hein et al., 1996; Allen et Steel, 2001; Bordewich, Semple et Talbot, 2004). Ces problèmes de réconciliation d'arbres sont connus pour être très difficiles à résoudre, dus à leur complexité algorithmique (NP-difficile) (Hein et al., 1996), et certains auteurs ont proposé différentes méthodes afin de réduire cette complexité. Le programme RIATA-HGT propose une méthode par décomposition qui détermine différentes régions de l'arbre qui peuvent être résolues indépendamment les unes des autres, en subdivisant l'arbre en différents sous-arbres (Nakhleh, Ruths et Wang, 2005). Le programme EEEP implémente également ce genre d'approche, mais permet aussi de restreindre la recherche pour différents types de SPR conduisant plus rapidement à de la discordance entre les arbres comparés (Beiko et Hamilton, 2006). Un autre algorithme, l'algorithme Prunier, utilise quant à lui de l'information sur la topologie, le support statistique et les longueurs de branches afin de procéder à la réconciliation (Abby et al., 2010).

D'autres méthodes utilisant les réseaux réticulés et les graphes permettent également la détection de transferts horizontaux (Makarenkov, 2001). Des méthodes utilisant finalement des métriques de distances telles que les bipartitions de dissimilarité (Boc, Philippe et Makarenkov, 2010) ou les distances Robinson et Foulds (Robinson et Foulds, 1981) peuvent également être en mesure de détecter raisonnablement les THG en plus d'établir la direction de ceux-ci.

1.5.8 Méthodes non-phylogénétiques

Avec l'amélioration des techniques de séquençage ainsi que la diminution de leur coût et l'augmentation de la vitesse des processeurs, les méthodes de détection de THG non-phylogénétiques ont vite été remplacées par leurs homologues, les méthodes phylogénétiques. Les méthodes de détection non-phylogénétiques étaient très utiles autrefois, lorsque le nombre de séquences disponibles était limité et que la résolution des arbres inférés était faible. Elles peuvent rester utiles cependant lorsqu'un grand nombre

de gènes sont à étudier. Elles sont effectivement très peu gourmandes en temps calcul et permettent d'éviter l'étape de reconstruction phylogénétique, qui elle devient rapidement très coûteuse en temps calcul plus la taille du jeu de données augmente (avec les méthodes probabilistes).

Les principales méthodes de détection non-phylogénétiques reposent généralement sur des mesures des caractéristiques génomiques de chaque espèce. L'une de ces méthodes, l'analyse de composition atypique (Poptsova, 2009), consiste à estimer le taux de G+C des différents gènes et si celui-ci est trop différent du taux de G+C du génome complet de l'organisme tel quel ou d'une espèce proche si le génome complet de l'espèce en question n'est pas disponible. Le taux de G+C a en effet tendance à être homogène au sein d'une espèce bactérienne (Daubin et Perriere, 2003) ainsi qu'à être similaire entre espèces proches (Ochman, 1996) et si une région génomique semble avoir une composition anormale face à l'ensemble du génome de cette espèce (ou d'une espèce étroitement apparentée venant le cas où le génome de cette espèce n'est pas disponible), celle-ci aura fort probablement été acquise lors d'un THG assez récent (Lawrence et Ochman, 1997). Ces méthodes sont cependant moins performantes si l'étude se fait exclusivement entre génomes d'espèces proches ayant un contenu génomique très similaire (Koski, Morton et Golding, 2001; Wang, 2001). De plus, ces méthodes auront plus de difficulté à détecter des THG anciens, puisque les gènes transférés auront eu le temps d'évoluer et d'obtenir les caractéristiques de l'organisme hôte (Lawrence et Ochman, 1997). D'autres méthodes paramétriques préfèrent comparer les contenus génomiques, lorsque les génomes complets des organismes sont disponibles. Certains auteurs ont bâti des méthodes de « naissance et mort » ou « birth and death » afin d'inférer l'histoire génomique des espèces et peuvent de cette façon supposer qu'une famille de gènes ait été ou non acquise de façon horizontale (Gu et Zhang, 2004; Hahn et al., 2005; Csuros, 2006; Iwasaki et Takagi, 2007; Cohen et Pupko, 2010). Les matrices de distances ont également été utilisées afin de vérifier la présence d'un gène ayant été acquis horizontalement (Kanhere et Vingron, 2009). Ceux-ci déterminent à l'aide d'une métrique de distance, la distance de Cook (Cook, 1979; Lorenz, 1987), la présence de séquences atypiques, obtenues en comparant la matrice de distance des séquences provenant d'un alignement de référence avec la matrice de

distances des alignements protéiques. Cette nouvelle méthode permet la découverte de THG entre espèces très éloignées et un certain nombre entre espèces proches.

La précision des approches phylogénétiques et non-phylogénétiques reste cependant encore très incertaine. Des résultats très différents les uns des autres ont été obtenus concernant la fréquence des transferts horizontaux et ceux-ci dépendent très fortement de la nature de la technique utilisée. Afin de ne pas s'avancer vers des pistes trop risquées, certains auteurs ont donc décidé d'éviter le plus possible le problème des THG en utilisant seulement des gènes ayant une très faible chance d'avoir été transférés. Très peu d'étude ont encore été faites sur l'impact d'un tel processus évolutif sur des données réelles ou simulées à une échelle phylogénomique, c'est-à-dire avec un grand nombre de gènes, et avec des méthodes robustes aux perturbations, telles que l'approche super-matrice. Ce sera l'un de nos sujets dans ce mémoire (voir chapitre 3).

1.5.9 Utilisation d'un noyau de gènes peu transférés

Afin d'essayer de limiter l'impact des THG, les études récentes sur les procaryotes utilisent des protéines qui ne sont pas beaucoup affectées par les THG (Brochier-Armanet et al., 2008; Elkins et al., 2008; Csuros et Miklos, 2009; Spang et al., 2010). Comme l'affirmait l'hypothèse de la complexité (voir section 1.5.1), certains gènes se trouvent moins affectés par les THG, parmi ceux-ci les protéines ribosomiques. La plupart des études récentes effectuées sur les Archées travaillent avec ces protéines (Brochier-Armanet et al., 2008; Elkins et al., 2008; Spang et al., 2010). D'autres études utilisent des gènes uniques conservés, ou universels, c'est-à-dire ayant un seul gène homologue par espèce étudiée (Daubin, Gouy et Perriere, 2002; Cox et al., 2008; Csuros et Miklos, 2009). Les phylogénies simple gène obtenues dans ces études sont compatibles entre elles et l'arbre des espèces résultant de l'utilisation de ces gènes, avec soit l'approche super-arbre ou super-matrice, est statistiquement mieux supportée que la phylogénie obtenue à l'aide de l'ARN ribosomique 16S (comparez figure 1.2 avec figure 1.7).

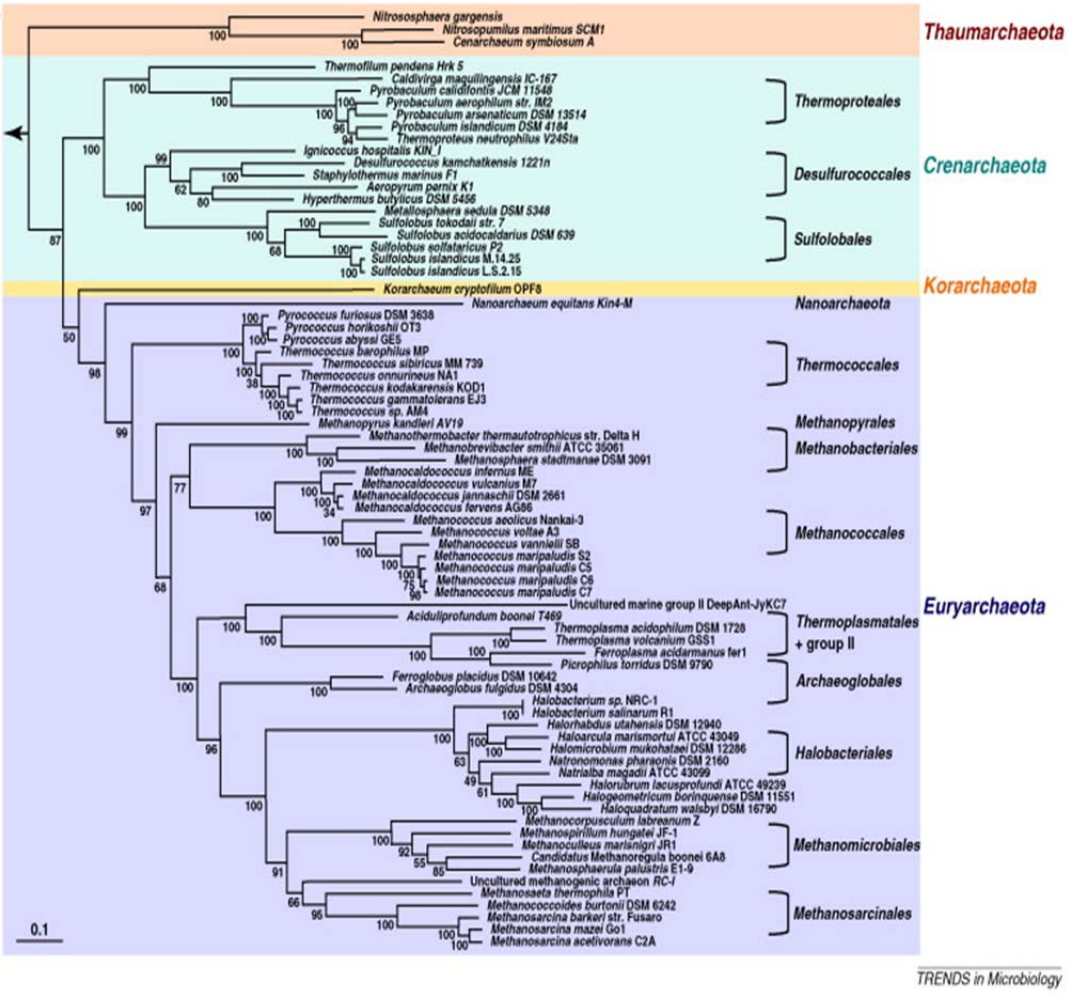


Figure 1.7 Arbre phylogénomique des Archées provenant de 53 protéines ribosomiques (107 séquences, 4683 positions). (Spang et al., 2010). Réimprimé de Trends in Microbiology, Vol. 18, Issue 8, Spang et al., Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota, pp 331-340, copyright 2010, avec la permission de Elsevier.

Hypothèses

Notre étude se concentre sur plusieurs points concernant la phylogénie et l'évolution des procaryotes, et principalement sur les Archées. Nos hypothèses reposent sur les modèles utilisés dans l'inférence des arbres phylogénomique ainsi que sur l'impact des THG sur différentes approches de reconstruction utilisées en

phylogénomique. Nous vérifions premièrement la phylogénie présentement acceptée chez les Archées et plus particulièrement les méthodes utilisées afin de reconstruire cette phylogénie. Deux études (Brochier-Armanet et al., 2008; Spang et al., 2010) proposent l'existence d'un troisième phylum d'Archée, les Thaumarchaeota, et stipulent que ceux-ci forment le groupe frère de tous les autres Archées. Nous faisons l'hypothèse que ce placement est artéfactuel dû à l'utilisation des Eucaryotes, un groupe externe très éloigné. Pour notre seconde hypothèse, nous envisageons que les transferts horizontaux de gènes (THG) aient moins d'impact sur les approches phylogénomiques lorsqu'un grand nombre de gènes est utilisé, puisque le taux de transmissions horizontales (ou THG) est minime comparativement au nombre de transmissions verticales. Nous devrions alors être en mesure de retrouver le signal dominant, que représente l'histoire verticale de l'évolution, malgré le fait qu'un certain nombre de gènes contiennent des transferts horizontaux.

Objectifs

Différentes manipulations phylogénétiques et protocoles de test sur les THG vont être élaborés afin de pouvoir confirmer ou infirmer nos différentes hypothèses. Le premier objectif sera alors de réanalyser le jeu de données plus récent mentionné précédemment (Spang et al., 2010) et de lui appliquer des modèles d'évolution de séquences plus complexes que ceux utilisés par les auteurs (CAT+GTR+Γ_4, CAT+GTR+Γ_4+Dayhoff6). Cette manipulation nous permettra de voir l'impact de ces modèles prenant en compte l'hétérogénéité substitutionnelle des différents sites, sensés éliminer une bonne part des artéfacts de reconstruction typiques, tels que les LBA. Le second objectif vise à voir l'impact des THG sur différentes approches phylogénomiques. À l'aide de simulations, nous comparerons deux approches super-arbres ainsi qu'une approche super-matrice par rapport à l'impact du taux de THG sur ceux-ci pour des jeux de données comprenant approximativement le même nombre de positions. Nous chercherons également à vérifier la puissance des approches phylogénomique pour différentes tailles de jeux de données.

Chapitre 2

Le premier article, qui sera très prochainement soumis à *Nature Reviews Microbiology*, a comme message principal que le choix du modèle d'évolution de séquences est très important lorsqu'on travaille à une très grande échelle taxonomique. Ce message est supporté par des études antérieures faites sur les métazoaires (Philippe et al., 2011a; Philippe et al., 2011b) qui utilisent les récents modèles d'évolution de séquences prenant en compte l'hétérogénéité du processus évolutif des sites d'un alignement, soient les modèles CAT et CAT+GTR (Lartillot et Philippe, 2004). Des comparaisons sont alors faites en réanalysant différents jeux de données publiés concernant la phylogénie des Archées. L'étude se focalise principalement sur l'article de Spang et al. (2010) qui utilise un modèle d'évolution de séquence considérant l'homogénéité des processus évolutifs pour tous les sites de l'alignement (par exemple les modèles JTT, LG, WAG or GTR). Les résultats obtenus dans notre article viennent contredire les positions de certains groupes publiées dans cet article et dans un autre article paru en 2008 (Brochier-Armanet et al.) qui soutiennent que les Thaumarchaeota sont groupe frère de toutes les Archées. Nos résultats soutiennent quant à eux que les Thaumarchaeota sont groupe frère des Crenarchaeota avec un bon support statistique (1 de probabilité postérieure). De plus, une seconde espèce, *Korarchaeum cryptofilum*, était difficile à positionner dans les deux articles parus traitant de ce sujet (Brochier-Armanet et al., 2008; Spang et al., 2010). Dans notre étude, cette espèce est désormais positionnée comme groupe frère des Crenarchaota et Thaumarchaeota avec un support de 0,99 en probabilité postérieure. Les résultats semblent être satisfaisants lorsqu'on considère que les modèles utilisés expliquent mieux les données que les modèles homogènes. Le faible nombre de taxons dans les différents groupes concernés ne peuvent cependant pas nous permettre d'affirmer que ces résultats représentent les positions définitives de ces groupes. Plus de taxons seront nécessaires afin de s'assurer de celles-ci.

Complex models of sequence evolution, the neglected component of prokaryotic phylogenetics

Jean-Christophe Grenier, Henner Brinkmann, Hervé Philippe

Centre Robert Cedergren, Département de Biochimie, Université de Montréal, Montréal, Québec, H3T 1J4, Canada

Corresponding author:

Hervé Philippe

Département de Biochimie, Université de Montréal, Succursale Centre-Ville, 2900 Boulevard Edouard-Montpetit, Montréal, Québec H3C 1J4, Canada

2.1 Abstract

The inference of ancient phylogenies is difficult. In the era of genomics, the two most important limitations are the combined confounding effects of horizontal gene transfers and of gene duplications, and the limitations of tree reconstruction methods. Since about ten years, research has focused on the first aspect in the case of prokaryotes, and on the second one in the case of eukaryotes. Here, we argue that it is time to take advantage of the largely improved tree reconstruction methods, in particular regarding the modelling of sequence evolution, for inferring the phylogeny of prokaryotes, which will be very helpful for the study of horizontal gene transfers.

Keywords: systematic error, phylogenomics, inconsistency, Archaea, compositional bias, site heterogeneity.

2.2 Questioning the tree of life

Phylogenetics of prokaryotes has been greatly influenced by the 1999 article of Ford Doolittle (Doolittle, 1999), in which the important impact of horizontal gene transfers (HGT) on the construction of the Tree of Life (ToL) was emphasized. In particular, Doolittle stated (p. 2124) "molecular phylogeneticists will have failed to find the "true tree", not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented by a tree". A heated debate about the existence of the ToL and the best way to represent the history of life (e.g. (Ochman, Lawrence et Groisman, 2000; Charlebois, Beiko et Ragan, 2003)) has been engaged among most of the prokaryotic phylogeneticists. HGTs were confirmed to play an important role during the history of prokaryotes, in particular, by allowing rapid adaptation to new environments (Valentine, 2007). However, the observation that genomes of closely related organisms may have a very different gene complement (e.g. only 39% of the genes shared by three strains of *Escherichia coli* (Welch et al., 2002)) does not demonstrate that horizontal inheritance outnumbers the vertical one, because most of these transferred genes are quickly lost (Hao et Golding, 2010). Quantitative estimates of the frequencies of HGTs, although still problematic in several respects, are consistently low, on average from a few up to ten per gene (Snel, Bork et Huynen, 2002; Mirkin et al., 2003; Beiko, Harlow et Ragan, 2005; Beiko et Hamilton, 2006; Dagan, Artzy-Randrup et Martin, 2008; Puigbo, Wolf et Koonin, 2010). Therefore, horizontal transmission of genes is million to billion times less frequent than vertical transmission (Philippe et Douady, 2003), thereby making the inference of the phylogeny (i.e., vertical history) the major task, since it represents by far the largest part of the history of prokaryotes. This does not mean that HGTs are not important and should not be inferred and should not be displayed as thin lines on the network of life, where the phylogeny of ToL will be the trunk.

Since 1999, researchers have pursued their efforts to infer the phylogeny of prokaryotes, but have been seriously disturbed by the question of HGTs. They took great care to one of the two crucial factors given by Doolittle ("they have chosen the wrong genes" (1999)) by selecting a core of rarely transferred genes, mainly encoding

ribosomal proteins (Brochier et al., 2002; Forterre, Brochier et Philippe, 2002; Matte-Tailliez et al., 2002; Brochier, Forterre et Gribaldo, 2004; Brochier et al., 2005; Ciccarelli et al., 2006; Brochier-Armanet et al., 2008; Elkins et al., 2008; Csuros et Miklos, 2009; Spang et al., 2010), because of the potentially misleading effect of HGTs on the inference of the species phylogeny (Box 2.1). The use of stringent criteria for the selection of orthologous genes led to relatively small datasets (e.g., 6,142 positions (Brochier-Armanet et al., 2008), 8,058 positions (Elkins et al., 2008), 5,222 positions (Foster, Cox et Embley, 2009) and 4,683 positions (Spang et al., 2010)), which are only one and a half to three times bigger than the concatenation of SSU+LSU rRNA (e.g., 3,305 positions (Brochier-Armanet et al., 2008)). The overall congruence among the phylogenies inferred from these various sets of rarely transferred proteins, from rRNA as well as from gene contents or oligonucleotide frequencies, argues in favour of our ability to recover the vertical signal of the history of prokaryotes (Snel, Bork et Huynen, 1999; Wolf et al., 2002; Philippe et Douady, 2003; Galtier et Daubin, 2008).

2.3 Phylogenetic reconstruction methods

With a few exceptions (Cox et al., 2008; Foster, Cox et Embley, 2009), researchers basically used the same phylogenetic inference methods that were available at the end of the 1990s (i.e. a site-homogeneous model based on a predefined matrix such as WAG (Whelan et Goldman, 2001) and a Γ distribution to handle rate across sites variation (Yang, 1994)). Thereby, they completely ignored the second major factor presented by Doolittle: “their methods are inadequate” (Doolittle, 1999). This is surprising, because many speciation events of the prokaryotic history are ancient (several billion years) and are expected to be very difficult to infer, even in the absence of HGTs. More precisely, systematic errors, which are due to violations of the underlying model of sequence evolution (in a probabilistic framework), are expected to play an important role, especially when a large number of genes are used.

In fact, for more shallow phylogenetic questions, the problem of systematic errors became the central aspect of phylogenomics already several years ago (Philippe et

al., 2005). For instance, the phylogeny of Saccharomycotina, a subgroup of ascomycetes, was shown to be influenced by the heterogeneity of the nucleotide composition across species (Phillips, Delsuc et Penny, 2004; Jeffroy et al., 2006). More generally, the position of fast evolving lineages, whose sequences are the most likely to violate the assumptions of the substitution model, is often inaccurately inferred (Box 2.2). For instance, in a phylogeny based on chloroplast genomes, the fast evolving monocots were inferred as the sister-group of all remaining angiosperms (Goremykin et al., 2003); the inclusion of slowly evolving monocots allowed to recover a topology in which *Amborella* is the sistergroup of all remaining angiosperms. The position of fast monocots is probably due toa systematic error, in this case a long-branch attraction artefact (LBA) (Felsenstein, 1978), subsequently alleviated by an enrichment of the taxonomic sampling (Soltis et al., 2004). Within animals, the positioning of the fast-evolving ctenophores and the fast-evolving acoels was similarly affected by long-branch attraction artefacts (Dunn et al., 2008; Hejnol et al., 2009) when a site-homogeneous model (Box 2.3) was used, i.e. they were attracted by the long branch of the outgroup or by the long branch of another fast-evolving ingroup (Box 2.2). In contrast, the use of a site-heterogeneous model (CAT or CAT-GTR (Lartillot et Philippe, 2004), Box 2.3) that better fits the data leads to a solution where the fast evolving animals are no longer grouped with other long branches (Philippe et al., 2007; Philippe et al., 2009; Philippe et al., 2011a; Philippe et al., 2011b).

Systematic errors play an important role in phylogenetic questions that are up to 10 times more recent than the phylogeny of prokaryotes. They should therefore affect more strongly the inference of prokaryotic phylogeny. We argue that reducing the impact of systematic errors, in particular through the use of complex models of sequence evolution, should become the central concern of phylogeneticists, who attempt to accurately infer the phylogeny of prokaryotes.

2.4 Systematic errors and archaeal phylogeny

We will illustrate the effect of systematic errors in the case of the phylogeny of Archaea. The first marine environmental rRNA sequences from mesophilic Archaea form two groups, one sister to the Crenarchaeota, the other located within Euryarchaeota (DeLong, 1992). The first group was named "mesophilic crenarchaeota" and contained species that are both abundant and that play a key ecological role via their ability of oxidizing ammonia (Karner, DeLong et Karl, 2001). However, based on the analysis of the complete and draft genome sequences of *Cenarchaeum symbiosium* (Brochier-Armanet et al., 2008), *Nitrososphaera gargensis* and *Nitrosopumilus maritimus* (Spang et al., 2010), a new phylogenetic position of "mesophilic crenarchaeota" as a sister-group to all remaining Archaea was obtained. Based on this result and some genomic considerations, it was proposed to consider this clade as a new archaeal phylum, named Thaumarchaeota. The underlying phylogeny was based on the analysis of 53 ribosomal proteins using simple, site homogeneous models of evolution, JTT+Γ (Brochier-Armanet et al., 2008) and LG+Γ (Spang et al., 2010).

These models assume that the substitution process is the same for all the positions of the alignment (Box 2.3), tuned by an amino acid exchangeability matrix (JTT (Jones, Taylor et Thornton, 1992) or LG (Le et Gascuel, 2008)), and that the only heterogeneity across sites corresponds to the rate of substitution. However, the hypothesis of homogeneity across sites is an obvious over-simplification of our biochemical knowledge (Miyamoto et Fitch, 1996; Halpern et Bruno, 1998; Lartillot et Philippe, 2004): for instance, some positions can only accept a negatively charged amino acid, others an aromatic one and others a small one. Models handling the heterogeneities of the substitution process across sites have been developed (Goldman, Thorne et Jones, 1998; Halpern et Bruno, 1998; Koshi et Goldstein, 1998; Lartillot et Philippe, 2004; Wang et al., 2008) (Box 2.3). In particular, the CAT and CAT-GTR models (Lartillot et Philippe, 2004) make use of the Dirichlet process to estimate from the data the distribution over sites of amino-acid propensities. In effect, Dirichlet processes are mixtures with an infinite number of categories, although the number of categories

represented in the sample will always be finite, and will be estimated from the data. These site-heterogenous models fit the data better than standard site-homogeneous models (Lartillot et Philippe, 2004; Lartillot, Brinkmann et Philippe, 2007; Philippe et al., 2007; Lartillot et Philippe, 2008; Lartillot, Lepage et Blanquart, 2009; Philippe et al., 2009; Sperling, Peterson et Pisani, 2009; Philippe et al., 2011b; Rota-Stabelli et al., 2011). More importantly, they detect more efficiently multiple substitutions (Lartillot, Brinkmann et Philippe, 2007), and therefore reduce the impact of systematic errors, such as the LBA artefact (Lartillot, Brinkmann et Philippe, 2007; Philippe et al., 2011b; Rota-Stabelli et al., 2011).

Table 2.1 Model fit (log likelihoods) estimated by cross-validation made with Phylobayes (Lartillot, Lepage et Blanquart, 2009).

Model/Alignments	Brochier-Armanet et al. (2008)	Spang et al. (2010)
JTT+Γ	0	0
LG+Γ	509±42	546±49
GTR+Γ	621±62	683±66
CAT+Γ	791±73	1152±145
CAT-GTR+Γ	1781±79	2046±114

The use of the site-heterogeneous CAT and CAT-GTR models in the case of Thaumarchaeota is *a priori* interesting. The outgroup is composed of the very distantly related eukaryotes, which represent a long branch that can easily attract any fast evolving archaeal group (Philippe et Laurent, 1998) (Box 2.2). We make the assumption that the observed position of Thaumarchaeota as the sister-group to all remaining Archaea is due to an LBA artefact, caused by violations of the over-simplistic models used. We first evaluated, using *cross-validation* (Lartillot et Philippe, 2008), the fit of site-homogeneous and site-heterogeneous models to the alignments of Brochier-Armanet et al. (Brochier-Armanet et al., 2008) and Spang et al. (Spang et al., 2010). In both cases, the ranking of the models (Table 2.1) was the same: all the site-homogeneous (WAG<LG<GTR) models fit less well the data than the site-heterogeneous models (CAT<CAT-GTR), in agreement with previous studies (Lartillot

et Philippe, 2004; Lartillot, Brinkmann et Philippe, 2007; Philippe et al., 2007; Lartillot et Philippe, 2008; Lartillot, Lepage et Blanquart, 2009; Philippe et al., 2009; Sperling, Peterson et Pisani, 2009; Philippe et al., 2011b; Rota-Stabelli et al., 2011).

The phylogeny was thus inferred with the best fit CAT-GTR model for both the alignments of Brochier-Armanet et al. (2008) (Fig. S2.1) and of Spang et al. (2010) (Fig. 2.1). The two trees were very similar, despite some differences in taxon sampling and therefore only the reanalysis of Spang et al. will be discussed. Although the phylogenies inferred with the LG+Γ and CAT-GTR+Γ models were mostly congruent, they differed in their deepest branching patterns. First, instead of being the sistergroup of all Archaea (Spang et al., 2010), Thaumarchaeota are the sistergroup of Crenarchaeota (Fig. 2.1). This grouping is congruent with rRNA based phylogeny (DeLong, 1992), and with the hypothesis that their alternative position as sistergroup of all Archaea is the result of an LBA artefact, generated by the use of an oversimplified model. In fact, a comparison of the results of Brochier-Armanet et al. (2008) and Spang et al. (2010) also supports the LBA hypothesis. Spang et al. use a better model (LG instead of WAG, Table 2.1) and a richer taxon sampling, two approaches known to reduce the misleading effect of LBA (Soltis et al., 2004; Philippe et al., 2005). Accordingly, they obtained a lower bootstrap support (89%) for Thaumarchaeota as the sistergroup of all Archaea than in Brochier-Armanet et al. (99%). These limited improvements only result in a small reduction of the impact of the systematic error; in contrast, the use of the site-heterogeneous CAT-GTR model allows to overcome the LBA artefact for both the species-poor and the species-rich alignments (see also (Philippe et al., 2011a)). Two other topological changes were observed (our Fig. 2.1 versus the Fig. 1 of Spang et al.) and were congruent with a lower sensitivity to the LBA artefact of the CAT-GTR model. The long branched *Ignicoccus* was no longer the sistergroup of all remaining Desulfurococcales, but rather of *Aeropyrum*+*Hyperthermus* within Desulfurococcales. Finally, the long branched *Korarcheum* became the sister-group of Thaumarchaeota+Crenarchaeota, instead of Euryarchaeota. Please note that, even as a sister-group to Crenarchaeota, Thaumarchaeota might be considered as an archaeal phylum, given their distant relationship to both Crenarchaeota and Euryarchaeota and their genomic distinctiveness

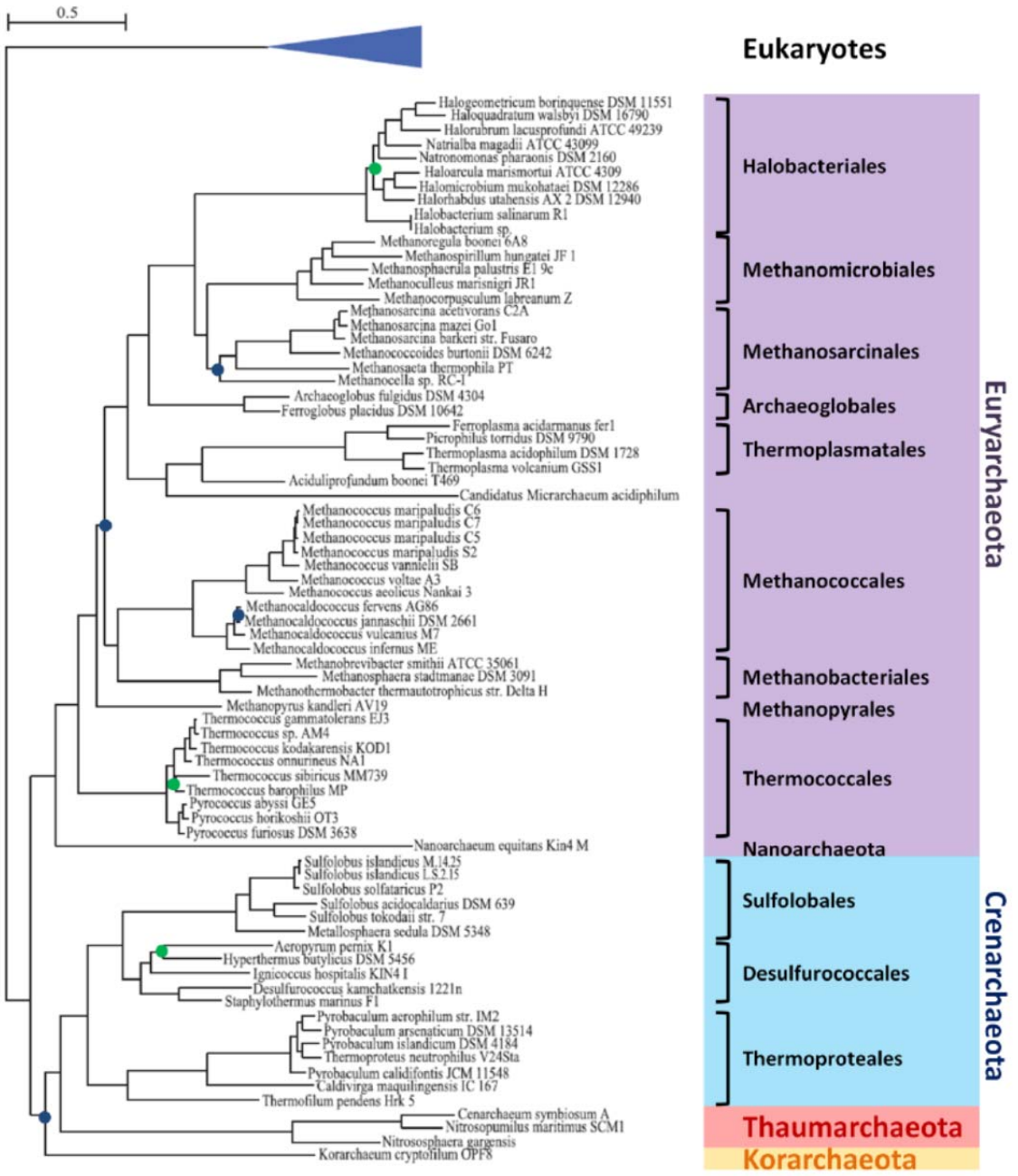


Figure 2.1 Archaeal phylogeny based on the 53 ribosomal protein dataset from Spang et al. (2010), which consists of 107 species (76 Archaea and 31 eukaryotes used as outgroup) and 4683 amino acid positions. The tree was inferred by PhyloBayes (Lartillot, Lepage et Blanquart, 2009) with the CATGTR+Γ_4 model of sequence evolution. Only posterior probabilities (PP) lower than 1.0 are indicated by coloured dots: blue PP ≥ 0.95 and green PP < 0.95. A cross-validation using a 80 species subset (Table 1) showed that the CATGTR+Γ_4 model was significantly better (-1499±878) than the LG+Γ_4 model used in Spang *et al.* (Spang et al., 2010).

noted by Brochier-Armanet et al. (2008) and Spang et al. (2010).

2.5 Compositional bias, another cause of reconstruction artefacts

Numerous other model violations are known (Philippe et al., 2005). For instance, most models are time-homogeneous, i.e. they assume that the amino acid (or nucleotide) composition is homogeneous across taxa. This hypothesis is often violated (e.g. (Lockhart et al., 1992; Delsuc, Phillips et Penny, 2003; Phillips, Delsuc et Penny, 2004)). Non- homogeneous amino acid composition is particularly likely to occur in Archaea that thrive in extreme environments (hyperthermophile, halophile, acidophile, etc.), because these environments impose constraints on the amino acid composition of the proteome (e.g. (Kennedy et al., 2001; Boussau et al., 2008)). As expected, a principal component analysis of the ribosomal protein alignments of Brochier-Armanet et al. (2008) and Spang et al. (2010) (Fig. S2.2 and S2.3) confirms the marked heterogeneity of amino acid composition across archaeal species. In particular, Halobacteriales, *Nanoarcheum* and *Methanopyrus* have very divergent compositions. Several non-stationary methods (Galtier et Gouy, 1995; Yang et Roberts, 1995; Foster, 2004; Blanquart et Lartillot, 2006) have been developed, and only one model (CAT-BP) (Blanquart et Lartillot, 2008) jointly models heterogeneity of the substitution process across sites and across branches. Unfortunately, the CAT-BP model is very time-consuming and the Markov chains did not converge with the large datasets we were reanalyzing. To address the large compositional heterogeneities of the alignments while still using a site-heterogeneous model, we used the Dayhoff recoding (Hrdy et al., 2004; Rodriguez-Ezpeleta et al., 2007b) which is an extension to amino acids of the well-known RY coding of nucleotides, for instance applied 20 years ago to position *Archaeoglobus fulgidus* (Woese et al., 1991). The phylogenies inferred from Dayhoff recoded alignments with the CAT-GTR model (Fig. S2.4 and Fig. 2.2) mainly differ for the position of the more compositionally biased species. Halobacteriales are sister-group of Methanomicrobiales (Fig. 2.2) instead of Methanomicrobiales+ Methanosarcinales (Fig. 2.1). More interestingly, *Nanoarcheum* is now the sistergroup of Thermococcales, in agreement with a careful genomic analysis (Brochier et al., 2005), but never

recovered in multigene phylogenies (Brochier, Forterre et Gribaldo, 2005; Gribaldo et Brochier-Armanet, 2006; Brochier-Armanet et al., 2008; Elkins et al., 2008; Spang et al., 2010);

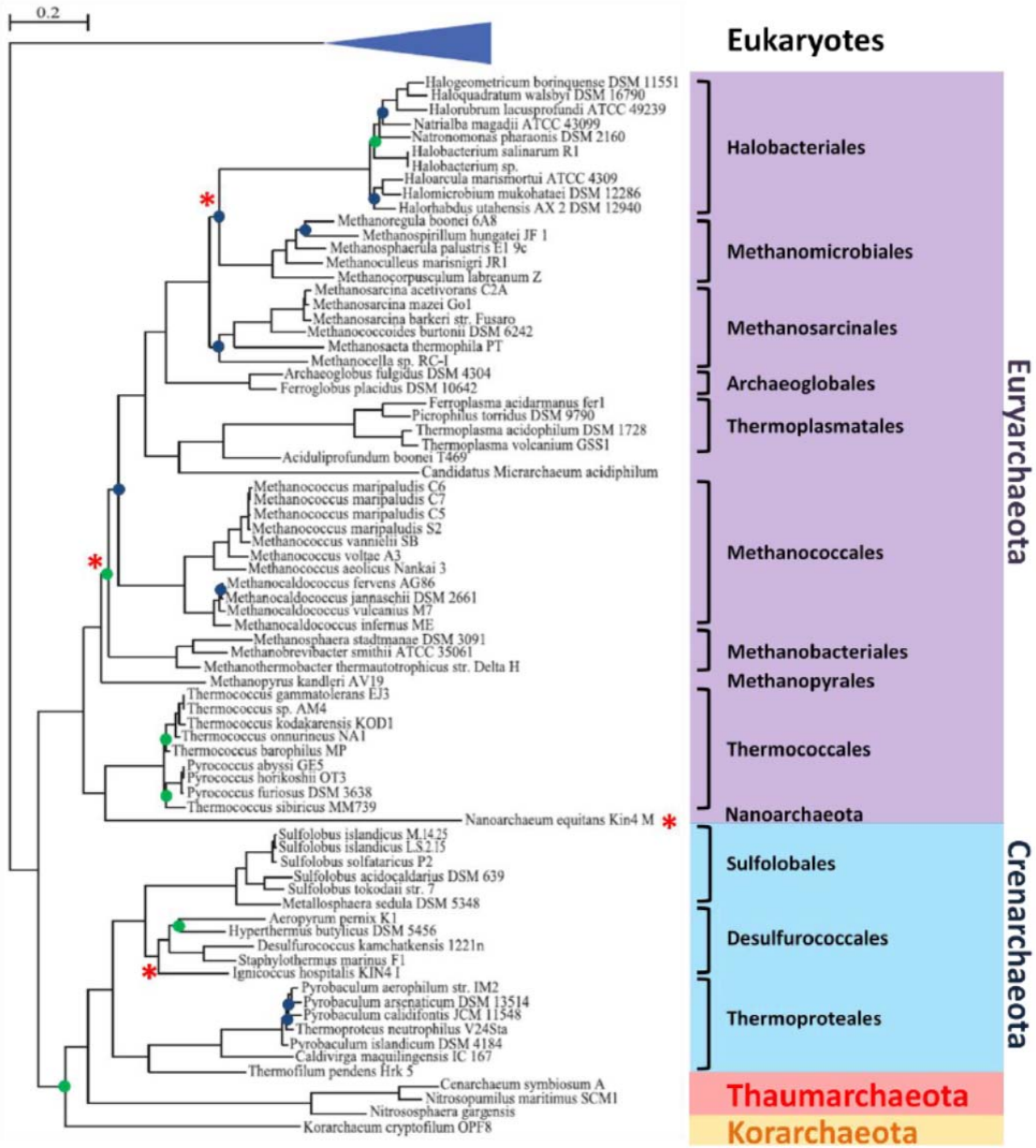


Figure 2.2 Phylogenetic tree inferred by PhyloBayes (Lartillot, Lepage et Blanquart, 2009) under the CATGTR+Γ_4 model based on the dataset of Fig. 2.1, which was recoded into the six Dayhoff functional categories (Hrdy et al., 2004). Only posterior probabilities (PP) lower than 1.0 are indicated by coloured dots: blue PP ≥ 0.95 and green PP < 0.95. Red stars indicate a topology that is different from the one shown in Fig. 2.1.

however, with the more limited taxon sampling of Brochier et al. (Fig. S2.4), in particular without *Korarcheum*, the placement of *Nanoarcheum* remains biased by an LBA artefact.

Doolittle accurately identified the two major limitations in inferring prokaryotic phylogeny (HGTs and reconstruction method), but he, and many scientists after him, put too much emphasis on HGTs to the detriment of inference methodology. Our reanalysis of two datasets (Brochier-Armanet et al., 2008; Spang et al., 2010) (Figs. 2.1, 2.2, S2.1 and S2.4) suggests that molecular phylogeneticists may have failed to find the correct phylogeny because "their methods are inadequate", and not because they have chosen the wrong genes. We have explored two important model violations, but many others have been identified and numerous model improvements have been developed (for review see (Philippe et al., 2005; Philippe et al., 2011b)). Moreover, other approaches to reduce systematic errors have been implemented, such as the removal of fast evolving positions (Brinkmann et Philippe, 1999; Pisani, 2004), the selection of genes (Brinkmann et al., 2005; Dopazo et Dopazo, 2005; Philippe, Lartillot et Brinkmann, 2005) or the removal of positions that violate model assumptions (Roure et Philippe, 2011). In conclusion, it is time for prokaryotic phylogenetics to take advantage of the major recent progresses capable to reduce the impact of the systematic error, along the lines initiated by the group of Martin Embley (Cox et al., 2008; Foster, Cox et Embley, 2009), and to simultaneously account for limitations introduced by HGTs and inference methods. In the long run, this will mean to develop models that simultaneously infer phylogeny, HGTs (Suchard, 2005; Bloomquist et Suchard, 2010), gene duplications, gene losses (Akerborg et al., 2009) and other genomic events using a complex and realistic model of sequence evolution. In the short run, this means to use the most accurate phylogenetic methods.

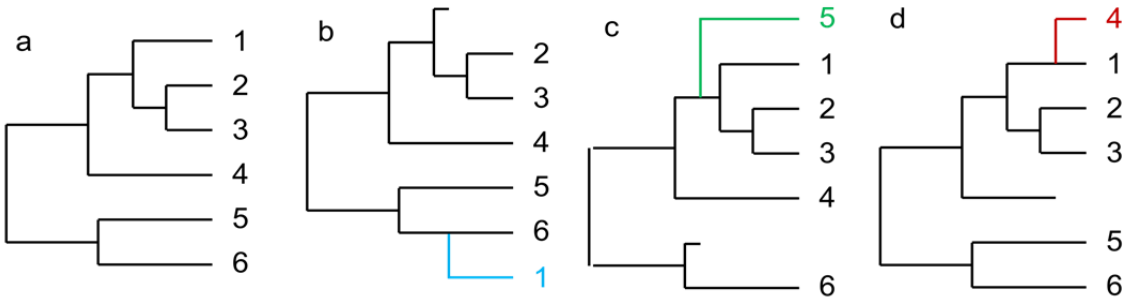
2.6 Acknowledgments

We wish to thank Nicolas Lartillot for computational resources. We also thank the Réseau Québécois de Calcul de Haute Performance for computational resources.

H.P. was supported by the Canadian Research Chair Program, H.P. and H.B. were supported by NSERC and J.C.G was supported by CIHR and Université de Montréal.

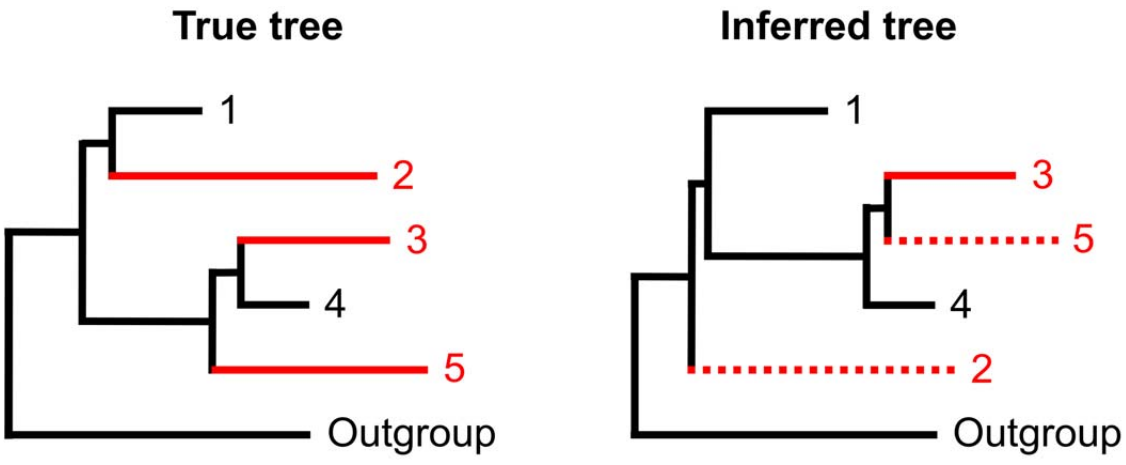
Box 2.1: The impact of horizontal gene transfer on phylogenetic inference

The horizontal transfer of a gene between distantly related taxa has a tremendous effect on the shape of the phylogeny. If, on the species phylogeny described in a, an ancestor of species 1 gets its gene from an ancestor of species 6, these two species will be closely related for this gene (b) while distantly related in the species phylogeny. The congruence among the three genes that independently underwent a single gene transfer (b-d) is very limited, only the grouping of species 2 and 3 being always recovered. The highly disturbing consequence of HGTs on tree topology has justified the use of genes that underwent none or very few HGT events. Fortunately, tens of genes, especially those that encode a protein that has numerous interactions (Cohen, Gophna et Pupko, 2011), appear to be almost free of HGTs (Puigbo, Wolf et Koonin, 2010) and therefore accurately reflect the species phylogeny. More interestingly, even the genes that have undergone several HGTs appear to have conserved a large amount of signal of the species phylogeny. Computer simulations have shown that the species tree inferred from tens of genes with up to tens of HGTs, using either a super-tree (Galtier, 2007) or a super-matrix approach (unpublished results), is almost identical to the correct species phylogeny, indicating the robustness of phylogenomics to HGTs.



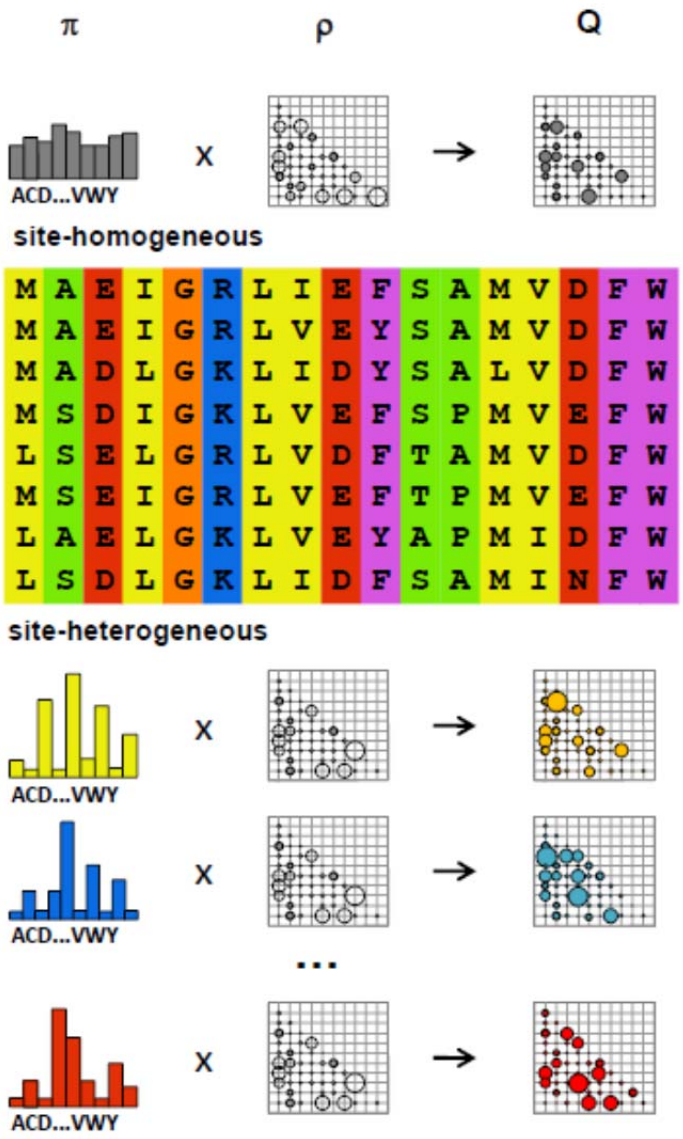
Box 2.2: Systematic errors and phylogenetic artefacts

The long-branch attraction (LBA) artefact, initially described for maximum parsimony (Felsenstein, 1978) and non-corrected distance methods, is probably the most common phylogenetic artefact. As shown in Fig. a, fast-evolving ingroup species (a long branch) can be erroneously attracted by another long branch, either the outgroup (for species 2), especially when distantly related (Philippe et Laurent, 1998) or another fast-evolving ingroup species (for species 5) are involved. Probabilistic methods, maximum likelihood and Bayesian inference, are expected to be more robust than maximum parsimony to LBA, because they take into account the branch lengths (Felsenstein, 1981). However, they are not immune to model violations. Long branches sometimes correspond to distant relationships, but often to high rates of evolution. In the latter case, the substitution process is expected to display anomalous statistical properties. For instance, an accelerated evolutionary rate can be achieved by an increased rate at variable positions, which will be correctly handled by classical probabilistic models, or by an increase of the number of variable positions (Germot et Philippe, 1999; Philippe, 2000; Gruenheit et al., 2008), which will not. Compositional bias, especially if it is driven by mutation, will more easily accumulate along long branches, on which purifying selection may be weaker. In addition, the nature of amino acids that are acceptable at a given position is more likely to change along these long branches (Roure et Philippe, 2011). Because of these numerous model violations concentrated in fast-evolving sequences, long-branch attraction is still frequently observed in probabilistic inferences.



Box 2.3: Site-homogeneous and site-heterogeneous models of sequence evolution

The most widely used models of sequence evolution for amino acids (e.g., JTT (Jones, Taylor et Thornton, 1992), WAG (Whelan et Goldman, 2001), or GTR (Lanave et al., 1984)) make the simplifying assumption that the evolutionary process is homogeneous across sites. The probability of substitutions of an amino acid by another one (the substitution matrix Q) is obtained by the product of the vector of stationary amino acid frequencies (π) by the matrix of instantaneous exchangeability rates (ρ). As a result, the probability of replacing a glutamate by an aspartate, is the same for any position of the entire alignment, irrespectively of its context within the protein structure. However, there is an entire spectrum of context dependant probabilities, for a substitution at a given position, which can go from almost no constraints, to hydrophobic or hydrophilic, to small or bulky, to negative or positively charged, to largely invariant. Site-heterogeneous models will assign positions to different categories, for instance surface exposed or buried positions. Several models have different exchangeability matrices ρ(Goldman, Thorne et Jones, 1998; Koshi, Mindell et Goldstein, 1999; Le, Lartillot et Gascuel, 2008; Pagel et Meade, 2008; Wang et al., 2008), but the number of categories is limited because of the large number of parameters. To reduce the number of parameters, while keeping the most significant information concerning the functional constraints, it is possible to use the same exchangeability matrix for all positions, but different vectors of stationary amino acid frequencies (Fig. b). The CAT model (Lartillot et Philippe, 2004) assumes that all exchangeability rates are identical, which speeds up computations by 10-20 times, while the CAT-GTR model (Lartillot et Philippe, 2004) infers the exchangeability rates from the data. Although sites-heterogeneous models constitute a major improvement over site-homogeneous ones, they still make over-simplifying assumptions, such as the independence of the positions (Robinson et al., 2003; Rodrigue et al., 2005) or the homogeneity of the substitution process over time (Roure et Philippe, 2011).


π
ρ
Q
ACD...VWY
X
site-homogeneous
M A E I G R L I E F S A M V D F W
M A E I G R L V E Y S A M V D F W
M A D L G K L I D Y S A L V D F W
M S D I G K L V E F S P M V E F W
L S E L G R L V D F T A M V D F W
M S E I G R L V E F T P M V E F W
L A E L G K L V E Y A P M I D F W
L S D L G K L I D F S A M I N F W
site-heterogeneous
ACD...VWY
X
ACD...VWY
X
...
ACD...VWY
X

Glossary

Clade posterior probability (PP): The probability that a node is correct given the data in Bayesian phylogenetics, which is conditioned on the assumption that the model used in the analysis is correct.

Bootstrap support (BS): A statistical analysis used to test the reliability of the nodes in phylogenetic trees. The bootstrap value of a node corresponds to the proportion of times that the group of species is present in the set of trees constructed from the resampled datasets (same length as the original dataset) created by independent random sampling with replacement of positions of the original dataset.

Cross-validation tests: This is a statistical method that allows the estimation of the fit of probabilistic models to the data, thereby identifying the one with the best fit. More precisely, a part of the data (here 9/10) is used to estimate all the parameters of a given model, subsequently these parameters are applied to estimate the likelihood of the rest of the data (here the 1/10 non-overlapping part). This random partition into learning and test sets is repeated several times (here 10). The underlying assumption is that the better model has a higher predictive capacity for unknown data and will therefore obtain the better likelihood scores, which are considered to be significant if they are higher than 1.96 standard errors.

Dayhoff recoding: A recoding of the original 20 amino acids into 6 categories, essentially based on similar functional properties. Thereby reducing both the compositional heterogeneity and the saturation level of the data at the cost of a weakened phylogenetic signal.

Homoplasy: Identical character state at a given position due to multiple substitutions and not to common ancestry.

Horizontal gene transfer (HGT): The horizontal, lateral transfer of genetic material between two organisms, in contrast to the standard patterns of vertical inheritance.

Hyperthermophiles: Organisms that have optimal growth temperatures of at least 80°C.

Long-branch attraction (LBA) artefact: An artificial grouping of long branched lineages due to their high evolutionary rates or to their long independent evolutionary history, will artefactually group the fast evolving sequences/species.

Monophyletic group (clade): A group of organisms or sequences that includes the most common ancestor and all of its descendants.

Systematic error: Is the consequence of the fact that certain properties of the data are violating assumptions of the model of sequence evolution used for the phylogenetic analyses. This kind of error is usually the stronger the more data are used in the analysis.

Supplementary Figures

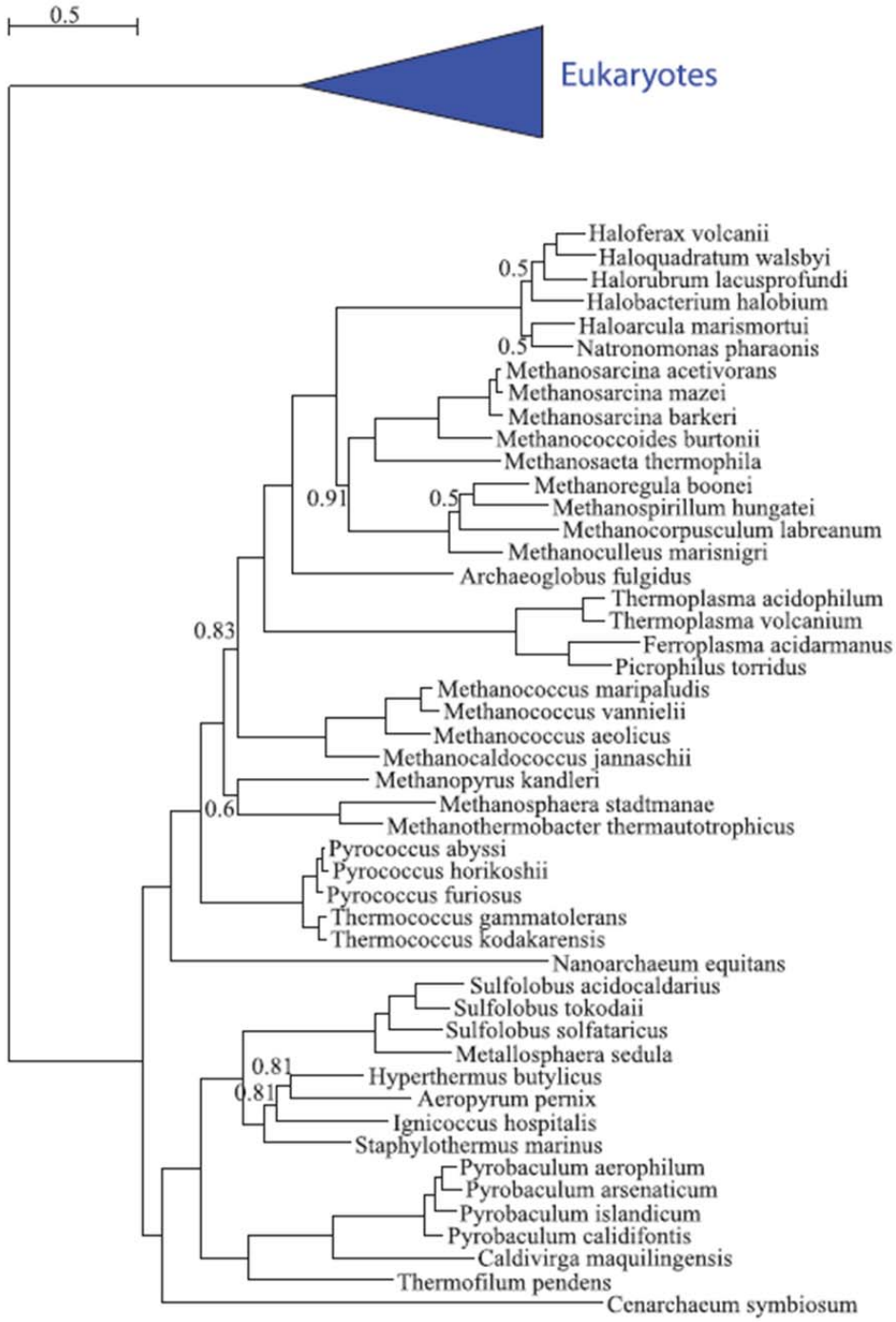


Figure S2.1 Archaeal phylogeny based on the 53 ribosomal protein dataset from Brochier-Armanet et al. (2008), which consists of 64 species (48 Archaea and 16 eukaryotes used as outgroup) and 6,142 amino acid positions. The tree was inferred by PhyloBayes with the CATGTR+Γ4 model of sequence evolution. Only PP-values for nodes that are supported by less than 1 are indicated.

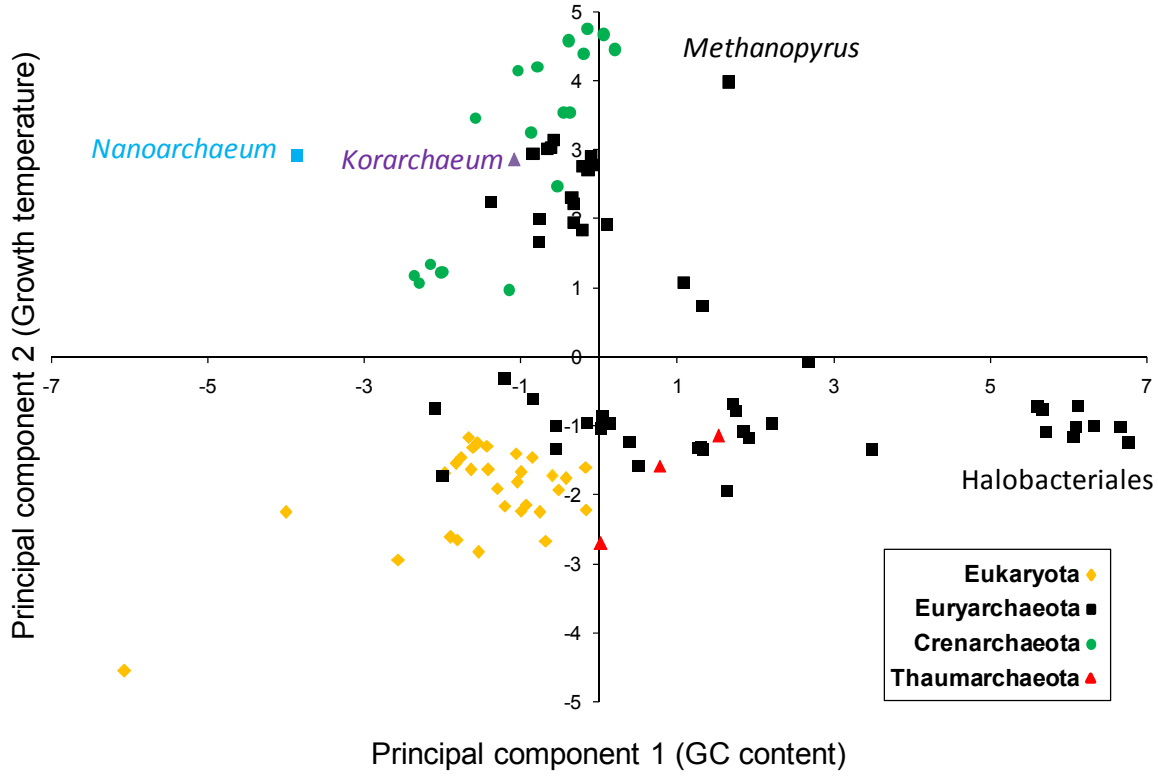


Figure S2.2 Principal component analysis of the 20 amino acid frequencies showing the heterogeneity of the composition among the 107 species of the Spang *et al.* (2010) dataset. The two first axes explain together 55.5 % of the variance and could be explained by the GC content of the sequences as well as the optimal growth temperature of the species (Kreil et Ouzounis, 2001). The further a given dot is located from the centre, the more the composition of the corresponding species is different from the mean value off all taxa. Therefore, the taxa that are the most distant from the centre have the most extreme compositional bias.

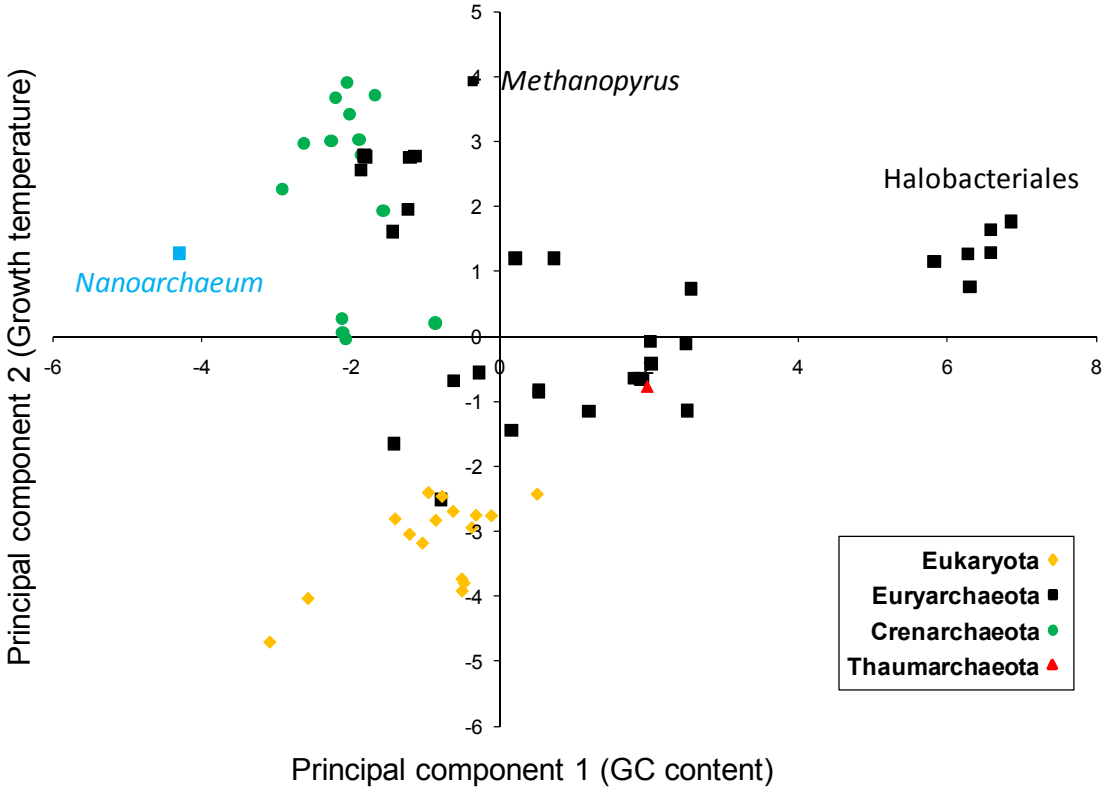


Figure S2.3 Principal component analysis of the 20 amino acid frequencies showing the heterogeneity of the composition among the 64 species of the Brochier-Armanet *et al.* (2008) dataset. The two first axes explain together 61.2 % of the variance and could be explained by the GC content of the sequences as well as the optimal growth temperature of the species (Kreil et Ouzounis, 2001).

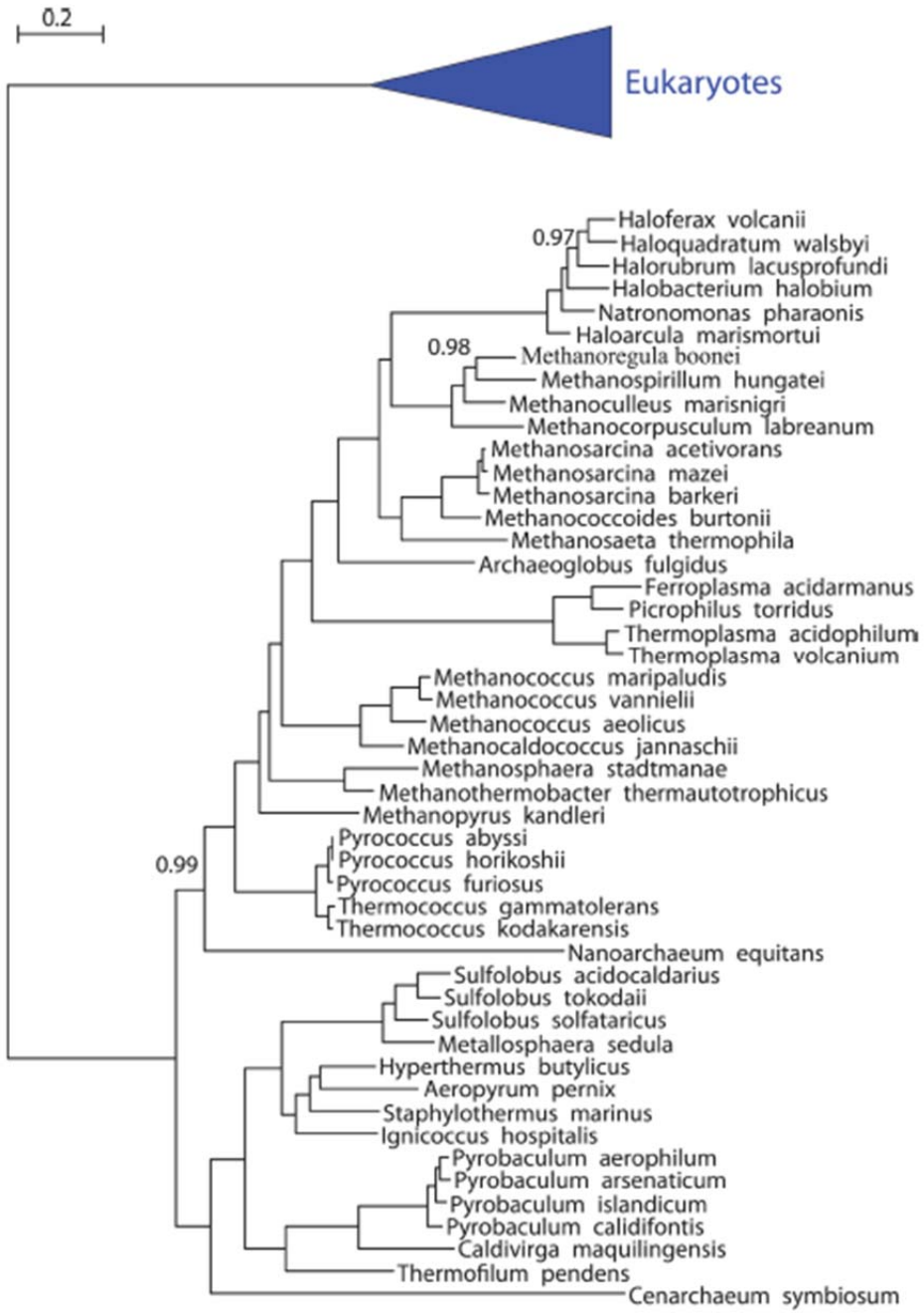


Figure S2.4 Archaeal phylogeny based on the 53 ribosomal protein dataset from Brochier-Armanet et al. (2008), which consists of 64 species (48 Archaea and 16 eukaryotes used as outgroup) and 6,142 amino acid positions. A Dayhoff recoding of the data into six functional categories (Hrdy et al. 2004) was performed and subsequently analysed with PhyloBayes under the CATGTR+Γ_4 model of sequence evolution. Only PP-values lower than 1 are given on the left of corresponding nodes.

Chapitre 3

La seconde étude se concentre à estimer l'effet sur l'inférence de la phylogénie des espèces d'un phénomène connu pour avoir un impact dans l'évolution et dans la transmission des gènes chez les procaryotes. Il s'agit des transferts horizontaux de gènes (THG). Les conséquences de ceux-ci sur différentes approches phylogénomiques sont étudiées à l'aide de simulations. Ces simulations de THG sont effectuées de façon aléatoire le long du temps ainsi qu'en choisissant aléatoirement les partenaires impliqués en considérant le nombre de branches disponibles à ce point précis dans le temps. Les approches de super-arbre et de super-matrice sont testées avec des jeux de données de différentes tailles, afin de voir l'impact de ce paramètre conjointement avec le taux d'évènements de transferts horizontaux. La conclusion principale de cet article est que l'approche super-matrice obtient généralement les meilleurs résultats lors de la reconstruction des arbres phylogénétiques et que ceux-ci sont de plus en plus précis lorsqu'on travaille avec un grand nombre de gènes. Cette étude est soumise au journal *Molecular Biology and Evolution*.

Do horizontal gene transfer events limit the accuracy of phylogenomics?

Jean-Christophe Grenier[1], Marie-Ka Tilak[2], Henner Brinkmann[1], Hervé Philippe[1]

[1]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal Québec, Canada. [2] Institut des Sciences de l'Evolution, UMR 5554 CNRS, Université Montpellier II.

[*]**Corresponding author**: Hervé Philippe

Département de Biochimie, Université de Montréal, C.P. 6128, Succursale Centre-ville, 2900, Bould. Edouard-Montpetit, Montréal, Québec, Canada H3C 3J7

Phone: (514) 343-6720

Fax: (514) 343-2210

Keywords: simulations, super-tree, super-matrix, tree of life, phylogeny

Running head: Horizontal gene transfers and phylogenomics

3.1 Abstract

The noteworthy effect of horizontal gene transfers (HGT) on the species tree inference has led researchers to perform their analyses with a core of rarely transferred genes, such as ribosomal proteins. However, the impact of HGT events on the accuracy of phylogenomic inference, i.e. based on hundreds of genes, has not been studied. In this work, we used simulations to fill this gap. Although, the accuracy of the species tree inference decreases with an increasing rate of HGTs, it remains appropriate even at high rate of HGTs. For instance, 98% of the groups are correctly recovered, when 100 genes with a mean of 30 HGT events per gene and 50 taxa are simultaneously considered. Our simulations also show that the super-matrix approach is slightly more accurate than the super-tree approach. Our results therefore suggest that HGTs do not seriously decrease the accuracy of the phylogenomic inference and that the prokaryotic tree of life could be inferred from a much larger set of genes than the core of rarely transferred genes.

3.2 Introduction

Horizontal gene transfers (HGTs) play an important role in evolution, in particular facilitating rapid adaptation to new ecological environments. Their impact on phylogenetic inference and species classification remains controversial. In particular, based on the very different gene content in the genomes of closely related strains (Welch et al., 2002), several authors suggested that HGTs are so frequent that the very concept of phylogeny should be abandoned and replaced by a network (Hilario et Gogarten, 1993; Doolittle, 1999; Ochman, Lawrence et Groisman, 2000; Charlebois, Beiko et Ragan, 2003; Doolittle et al., 2003; Koonin et Wolf, 2008; Boucher et Bapteste, 2009). However, even if genes can be frequently acquired in an organism, their probability of fixation in the population is extremely low (Hao et Golding, 2006; Marri, Hao et Golding, 2006), i.e. the fate of a gene acquired by HGT is generally to be lost. Accordingly, current estimates of the frequency of HGTs (Snel, Bork et Huynen, 2002; Mirkin et al., 2003; Beiko, Harlow et Ragan, 2005; Dagan, Artzy-Randrup et Martin, 2008; Puigbo, Wolf et Koonin, 2010) are rather low, less than ten on average per gene family in Bacteria. In consequence, horizontal transmissions of genes are millions times

less frequent than vertical transmissions (Philippe et Douady, 2003). Since the trunk (the phylogeny) is millions times thicker than the lianas (the non-phylogeny part of the network), it is of great interest to accurately infer this phylogeny (the tree of life).

However, even if HGTs are relatively rare, their impact on the structure of the phylogeny can be tremendous (when donor and recipient are distantly related), and more importantly their effects are cumulative. It is therefore expected that reconstructing the phylogeny of prokaryotes will be seriously complicated by HGTs (Bapteste et al., 2009), all the more so that inferring deep relationships is already difficult without HGTs because of model violations (for a recent review see (Philippe et al., 2011b)). Thus far, phylogenetic inference has therefore been mainly based on a core of non- or rarely transferred genes (Brochier et al., 2002; Forterre, Brochier et Philippe, 2002; Matte-Tailliez et al., 2002; Brochier, Forterre et Gribaldo, 2004; Brochier et al., 2005; Ciccarelli et al., 2006; Brochier-Armanet et al., 2008; Elkins et al., 2008; Csuros et Miklos, 2009; Spang et al., 2010). These genes are often performing informational tasks (e.g. translation and transcription), and are usually part of major protein complexes, thereby limiting the probability of successful HGTs according to the complexity hypothesis (Jain, Rivera et Lake, 1999; Cohen et Pupko, 2010). However, they are not numerous and often small (e.g., ribosomal proteins), limiting the number of positions that can be used: 6,142 amino acid position (Brochier-Armanet et al., 2008), 8,058 (Elkins et al., 2008), 5,222 (Foster, Cox et Embley, 2009), or 4,683 (Spang et al., 2010). This is much less than used for phylogenomic studies in eukaryotes, which are generally based on >20,000 positions and usually do not lead to fully supported trees (e.g., (Rodriguez-Ezpeleta et al., 2005; Dunn et al., 2008; Philippe et al., 2009; Burki et al., 2010)). In addition to this potential lack of statistical power, informational proteins might have some specificities that could lead to reconstruction artifacts. Hence, it would be very helpful to use other genes, so as to (i) corroborate the phylogenetic structure inferred mainly from translational apparatus and to (ii) increase the phylogenetic resolution.

This approach would imply the use of numerous other prokaryotic genes, but genes that may be subject to more frequent horizontal transfers. Surprisingly, very few

studies have analyzed the accuracy of phylogenetic analyses based on multiple genes that have underwent HGTs. An empirical study (Brochier et al., 2002) has shown that the concatenation of genes with rare HGTs yields in a phylogeny very similar to the rRNA or ribosomal protein trees. The concatenation of genes with frequent HGTs (e.g., tRNA synthetases) results in a less similar phylogeny, which is nevertheless still quite comparable (Brochier et al., 2002). Galtier (Galtier, 2007) performed an interesting study using simulations allowing HGT events to occur randomly in both time and tree space. The simulated sequences were analyzed by a super-tree method, Matrix Representation with Parsimony (MRP) algorithm (Baum, 1992; Ragan, 1992). Even with a high frequency of HGTs (a mean of 12 HGT events per gene (for 100 genes), for a 40 species tree), the super-tree is highly accurate. 80% of the nodes of the reference tree were recovered, which is considerably higher than the level of congruency (37%) observed among single gene trees and their super-tree. Despite the use of a limited number of proteins (20), super-tree methods are therefore robust towards elevated levels of HGT, at least when they occur randomly (Galtier, 2007). Finally, the accuracy of a genome reconstruction method based on Blastp scores (Clarke et al., 2002) was estimated under different patterns of HGT (Beiko, Doolittle et Charlebois, 2008). Despite the interesting scenarios tested in this article, the limited accuracy of the method in the absence of HGT makes this approach less promising. Despite their great interest, these few studies did neither explore the widely used super-matrix approaches nor analyze datasets of phylogenomic size.

The aim of this study is to evaluate the accuracy of super-matrix and super-tree methods when a large number of genes are used and when HGTs are frequent (for an overview of the protocol, see fig. S3.1). We performed simulations with Galtier's method, but used up to 200 genes. Super-matrix, and to a lesser extent super-tree, methods appear to be robust to a high frequency of randomly distributed HGTs (up to 50 for a 50 species tree), suggesting that the genes affected by HGTs should also be considered when inferring the prokaryotic phylogeny using phylogenomics.

3.3 Materials and methods

HGT simulations

The program of Galtier (Galtier, 2007) was used to perform simulations, as briefly described below. The ultrametric reference tree containing 50 species is shown on supplementary figure S3.2, its shape being inferred from a multi-gene phylogeny of Archaea. Starting from this ultrametric tree, genomic rate changes (ρ) were introduced in the tree. Then HGT events (τ) are simulated using the SPR ("Subtree Pruning and Regrafting") moves to create the trees containing these changes. The HGTs are uniformly introduced on randomly selected branches all along the tree and receivers and donors are chosen from a Poisson law distribution. After the HGT step, gene specific rate change events (ρ') and various gene tree diameters are applied, thereby simulating gene-specific selection constraints. Once the gene trees obtained, protein sequences were simulated under the JTT model of sequence evolution (Jones, Taylor et Thornton, 1992) with rates across sites heterogeneity defined by a gamma distribution with a shape parameter of 0.5. We also modified the PhyloBayes program (Lartillot, Lepage et Blanquart, 2009) in order to simulate sequences under the site-heterogeneous CAT-GTR model (Lartillot et Philippe, 2004); the parameters of this model were estimated on an alignment of 47 archaeal proteins (see supplementary materials for details).

The values of all the parameters of the simulations in our and Galtier studies are shown in table S3.1. The main differences are the number of genes (20 in Galtier versus 5, 10, 25, 50, 100 and 200 here) and the number of HGTs (0, 1, 3, 6 and 12 in Galtier versus 0, 5, 10, 20, 30, 40, 50 and 100 here), for which we explored a much wider range of parameters. In particular, the number of genes considered here is similar to what is currently used in phylogenomics and it can easily be obtained when comparing prokaryotic genomes. The other parameters that were less explored than in (Galtier, 2007) have been shown to have a limited impact on the outcome. 100 simulated replicates were performed for data sets using less than 100 genes and 10 simulated replicates were used for datasets using 100 and 200 genes.

Phylogenetic inference

All tree inferences of simulated data were performed with RAxML (Stamatakis, 2006) under a LG+F+Γ_4 (Le et Gascuel, 2008) model of sequence evolution (empirical amino acid frequencies and four discrete gamma categories) with 100 bootstrap

replicates. The single gene alignments were concatenated into a super-matrix with SCaFoS (Roure, Rodriguez-Ezpeleta et Philippe, 2007). Super-trees were inferred with two different programs (i) SuperTriplets v. 0.31 (Ranwez, Criscuolo et Douzery, 2010) and (ii) Super Distance Matrix (SDM) (Criscuolo et al., 2006). SDM was used in concert with PhyD* (Criscuolo et Gascuel, 2008) to construct a super-tree using a NJ-like algorithm from the distance matrix created by SDM. SDM transforms the single gene trees into distance matrices, combines them into a single super-distance-matrix and infers a tree from this latter matrix. SuperTriplets uses triplets metric to minimize the

Table 3.1 Combinations of genes with different HGT rates, used to build datasets with different mean rates of HGT, which have been defined as heterogeneous datasets, whose results are compared to the ones with homogeneous rates (used in fig. 3.4).

#HGT	0	5	10	20	30	40	50
Average of 5 HGT	90	0	0	0	0	0	10
	65	0	20	15	0	0	0
	25	50	25	0	0	0	0
	75	0	0	25	0	0	0
	50	0	50	0	0	0	0
Average of 10 HGT	65	0	0	20	0	15	0
	80	0	0	0	0	0	20
	0	50	25	25	0	0	0
	25	0	50	25	0	0	0
	75	0	0	0	0	25	0
	50	0	0	50	0	0	0
Average of 20 HGT	0	0	50	25	0	25	0
	25	0	0	50	0	25	0
	60	0	0	0	0	0	40
	50	0	0	0	0	50	0
Average of 30 HGT	20	0	0	0	50	0	30
	30	0	0	10	0	20	40
	35	0	0	0	0	25	40
	40	0	0	0	0	0	60
	25	0	0	0	0	75	0

distance between all trees. It uses agglomeration construction and aims to do a triplet asymmetric median super-tree. Because of computing time constraints, the super-matrix

approach was performed on only 10 replicates, while the super-tree approach was performed on all 100 replicates.

In addition, since the genes of a genome display largely varying HGT rates, we created data sets with varying HGT rates per gene, by combining genes simulated under several homogeneous HGT rates. Twenty different combinations were generated, with on average 5, 10, 20 or 30 HGTs per gene (Table 3.1). These more realistic conditions were studied only with 100 genes and simulations done with ρ' equal to 20 and ρ equal to 0, because genome rate changes are simulation-specific.

Accuracy and congruency measurement

All the inferred tree topologies were compared to the reference tree used for the simulations based on their bipartitions, assuming binary trees with identical taxon sampling. Therefore, the taxon sampling of the reference tree has to be adapted in the case of single gene tree, because gene losses (between 0 and 33%) were allowed in the simulations. A bipartition is said to be congruent/incongruent when we compare bipartitions for an inferred gene tree with the reference tree (Robinson et Foulds, 1981), and accurate/inaccurate (or concordant/discordant, see (Beiko, Doolittle et Charlebois, 2008)) when we compare an inferred super-tree or super-matrix species tree with the reference tree. The topologies of the two sub-trees defined by a bipartition may be not identical.

3.4 Results

As expected, the congruence of the single gene trees with the reference tree (fig. 3.1) is rapidly decreasing when the mean number of HGTs per gene goes from 0 to 100. The congruence in the absence of HGTs has only a mean value of 88%. Stochastic error is likely the main reason for this low value, because the congruence for long genes is higher than for small ones (fig. S3.3). As the rate of HGTs increases, congruence decreases rapidly (59% with 5 HGTs and 46% with 10) and is asymptotically approaching zero, with less than 10% of congruence for the highest HGT rates (100

simulated HGT events per gene). Similar results are obtained with different values of ρ and ρ', genome and gene rate changes (data not shown). This is in agreement with the results of Galtier (2007) despite the use of a slightly different metrics and indicates that our simulated single gene trees are highly different from the reference species tree.

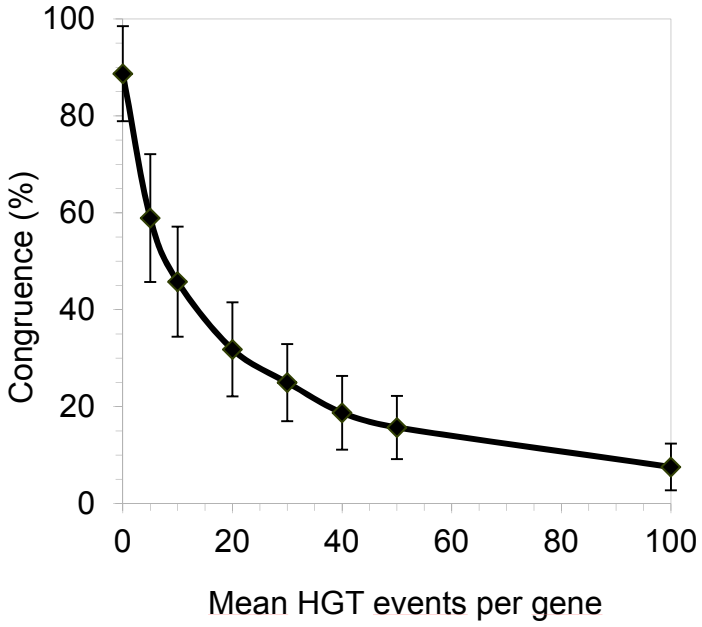


Figure 3.1 Simulation of single gene alignments with diverse levels of HGT events and estimates of the level of congruency of the inferred single gene trees with respect to the reference tree. The figure shows the results from the analysis of ten simulated genomes each with 100 genes and with 20 genomic rate change (ρ) and 20 gene-specific rate change events (ρ'). Sequences of 50 species were simulated and each single gene has a varying taxon sampling created by the random loss of 0-33% of the taxa.

In contrast, the accuracy is greatly enhanced when 100 genes are simultaneously used to infer the phylogeny (i.e., under phylogenomic conditions), both for the super-matrix, and the two super-tree approaches (fig. 3.2). Notably, the species tree inferred by super-matrix has a perfect accuracy up to a level of ten HGT events per gene (equaling a total of 1000 HGTs). A regular decrease of the accuracy for the three approaches under study is observed after 20 HGTs per gene, but the decline is slow with still 44% at 100 HGTs per gene (i.e., with 10,000 random HGT events). The super-matrix approach

slightly outperforms the SDM super-tree approach, which itself outperforms the SuperTriplets approach, which has an accuracy similar to the one of single genes with 100 HGTs (~10%). Different values of ρ and ρ', genome and gene rate changes have a negligible effect (fig. S3.4-3.6). These results demonstrate that phylogenomics is accurate when a realistic amount of random HGTs is assumed and model violations are

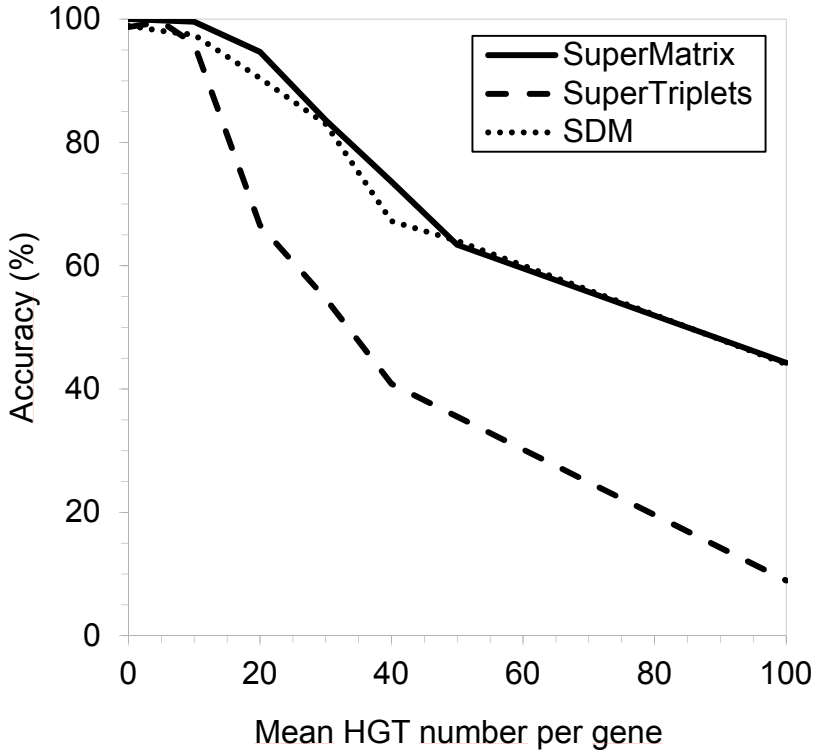


Figure 3.2 Phylogenomic approaches via the creation of a super-matrix or the super-tree approach based on simulation studies with diverse HGT levels. The conditions of the simulations were identical to the ones in Fig. 1 (ten replicates per data point and each data set contains 100 genes; ρ=20 and ρ'=20). Different methods were used to analyze the data, the super-matrices were analyzed with RAxML under the LG+F+Γ4 model and a consensus of the single gene phylogenies, also inferred by RAxML under the same model, was established with two super-tree approaches. The curves represent the accuracy, i.e. the percentage of nodes of the inferred species tree that are present in the reference tree, which was used for the simulations. The high accuracy values especially in the case of the super-matrix and the SDM approach, are demonstrating that phylogenomic approaches are much more resistant to high HGT levels than single genes.

minimal (i.e., sequences were simulated with the JTT+Γ model and inferred with the LG+Γ model), despite the fact that single gene trees are highly incongruent.

While keeping all other parameters identical to the ones of figure 3.2 (i.e., ρ and ρ' equal to 20), the accuracy is estimated for a number of genes that varies from 5 and 200 (fig. 3.3). As expected, the more HGT events, the greater the number of genes necessary to obtain the highest accuracy. In the absence of HGTs, an accuracy close to 100% is obtained with 20 genes. But, with 20 HGT events, the use of 200 genes is necessary to obtain a similar accuracy. The hierarchy of the three methods previously observed (super-matrix ≈ SDM > SuperTriplets) is globally recovered, in particular the poor performance of SuperTriplets. SDM appears to be less accurate when HGTs are rare (<20), but the most accurate when HGTs are frequent (100 per gene). The important message emerging from figure 3.3 is that the accuracy of the SDM and super-matrix approaches always increases when more genes are added, arguing in favor of the use of large datasets to infer the prokaryotic phylogeny.

Up to now the number of HGT events are homogeneous across genes (except for the stochastic variation of the simulation process), an unrealistic assumption, since it is well known that the numbers of HGTs per gene vary widely among genes (Puigbo, Wolf et Koonin, 2010). To simulate more realistic conditions, various combinations of heterogeneous rates of HGTs per gene were created (table 3.1), yielding 20 datasets containing 100 genes. Interestingly, at constant mean number of HGTs, the accuracy of the super-matrix method is enhanced under heterogeneous conditions with respect to homogeneous HGT rates (fig. 3.4). In particular, with 3000 HGTs (an average of 30 per gene), the average accuracy under heterogeneous rates is 99%, whereas it is only 88% under homogeneous rates. The results obtained in the two super-tree approaches were quite comparable (fig. S3.7-3.8).

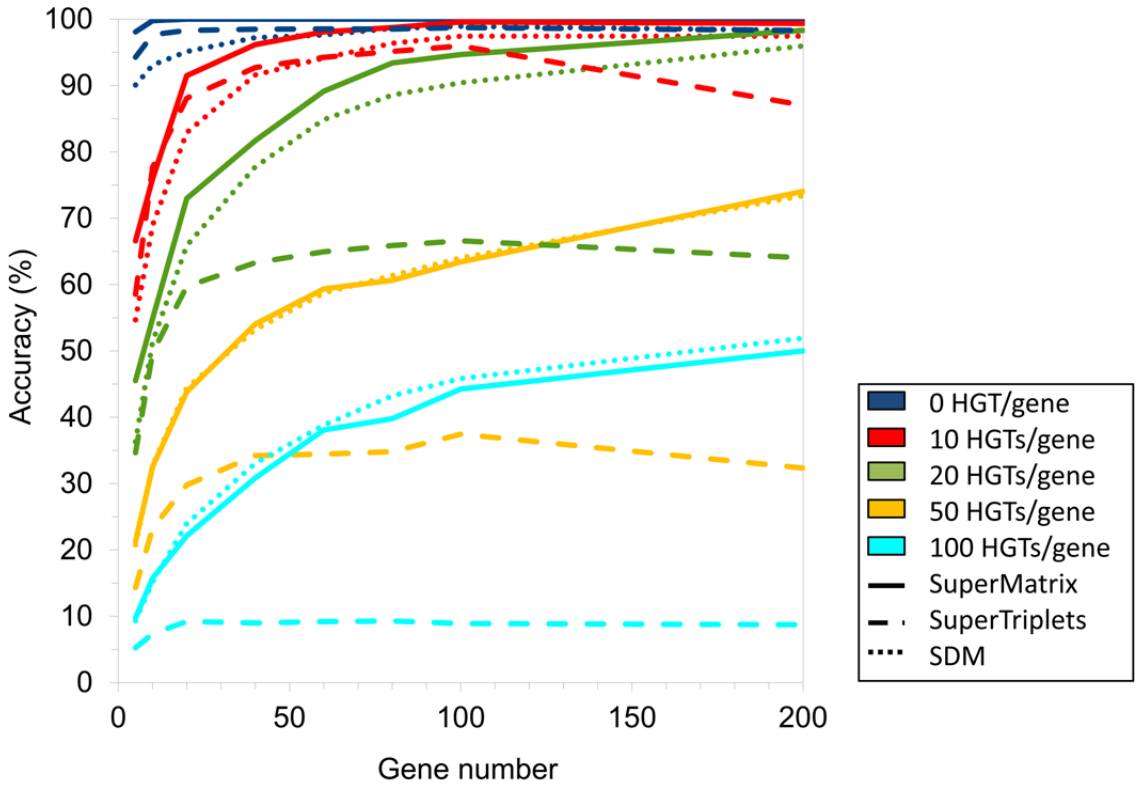


Figure 3.3 Estimation of the robustness of phylogenomic inference under diverse combinations of HGT rates and gene numbers. The curves depict the accuracy of the estimated species trees in comparison to the reference tree. For analyses with less than 100 genes, 10 replicates were analyzed for each data point in the super-matrix approach and 100 replicates in the two super-tree approaches, whereas for all datasets with 100 or more genes only 10 replicates were performed (ρ=20 and ρ'=20).

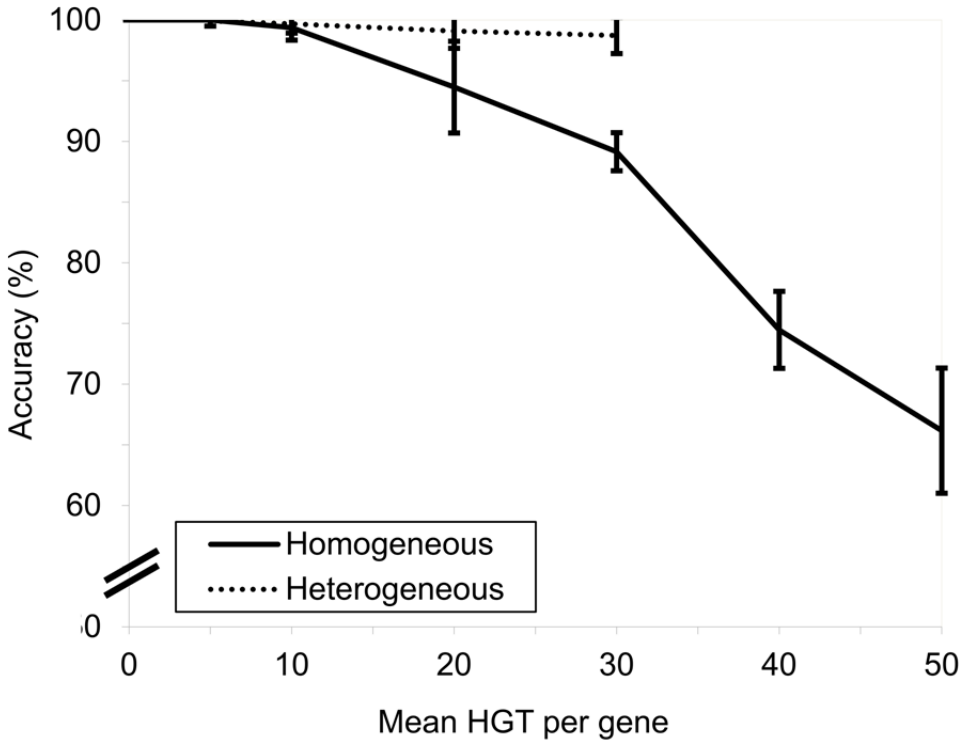


Figure 3.4 Comparison of the effect of the heterogeneity of HGTs rates across genes on the accuracy of the super-matrix approach. The curves represent the accuracy of the estimated species trees in comparison to the reference tree that was used to do the simulations. The simulation parameters were ρ=0 and ρ'=20 based on 10 replicates per data-point with each dataset consisting of 100 genes. The combinations of different HGT levels used in these analyses are described in table 1.

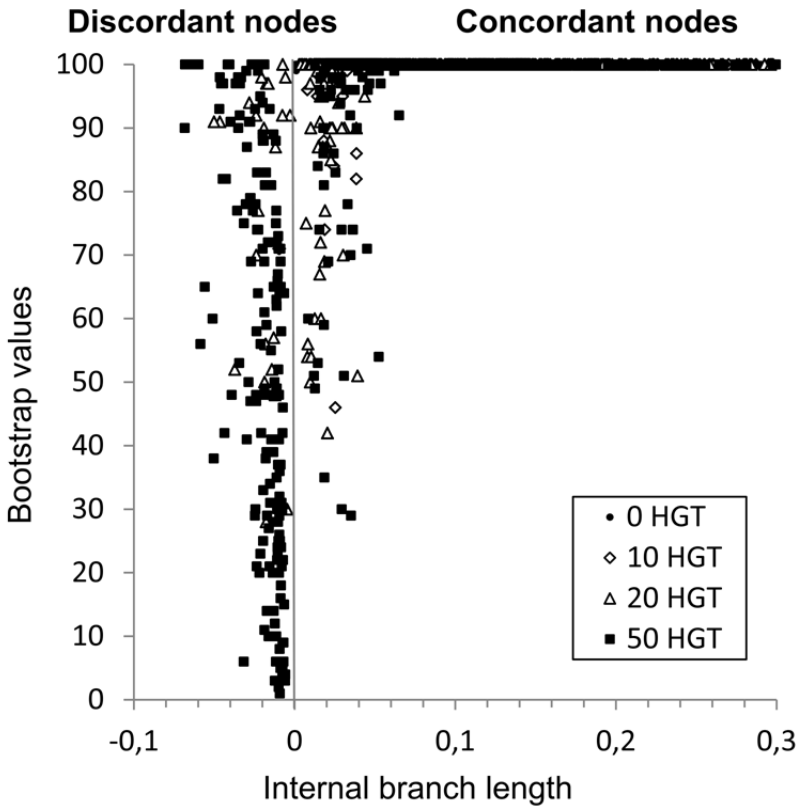


Figure 3.5 Exploring the correlation of the robustness of internal nodes measured in bootstrap support in comparison to the length of their basal branches. Internal branch lengths are classified according to their relation to the reference tree. Dots left to 0 (on the x axis) are inconsistent, i.e. the negative branch length represents a node that is not present in the reference tree. Dots right to 0 are consistent. Each dot represents 10 replicates; all of the HGT categories contain 47 nodes (470 dots per data point). All these results have been obtained with super-matrices based on 100 genes.

In order to better characterize how HGTs decrease the phylogenetic accuracy, the relationship between the length of an internal branch and its statistical support for the 100 genes super-matrix is shown in figure 3.5. Branch length and bootstrap proportions (BP) were computed for super-matrices at various levels of HGTs (0-10-20-50). When the bipartition is concordant with the reference tree, the dots are put on the right side of the figure, and when discordant on the left side (with negative branch length values for the aim of the display). As expected from previous results (fig. 3.2), the concordant branches represent the large majority (1,681 out of 1,880). Interestingly, 1,557 out of these 1,681 concordant branches have a 100 % bootstrap support. In contrast, almost all

of the discordant branches (199 in total; 25 for 20 and 172 for 50 HGTs per gene) are associated with small branch length (< 0.05), with most of them supported by low BP values (only 10 having a support of 100%). This result (fig. 3.5) is important for the analysis of real data, since it suggests that the rate of highly supported incorrect bipartitions (false positives) is small when large datasets with a high level of HGTs are considered, and that the main problem will be the lack of resolution.

3.5 Discussion

The fear that including genes that underwent HGTs will prevent an accurate phylogenetic reconstruction has drastically limited the number of studies of the phylogeny of prokaryotes, despite the availability of numerous complete genomes. To avoid the supposed highly negative effect of HGTs, the few phylogenetic studies either focused on a limited number of genes that are less likely to be transferred (Brochier et al., 2002; Forterre, Brochier et Philippe, 2002; Matte-Tailliez et al., 2002; Brochier, Forterre et Gribaldo, 2004; Brochier et al., 2005; Ciccarelli et al., 2006; Brochier-Armanet et al., 2008; Elkins et al., 2008; Csuros et Miklos, 2009; Spang et al., 2010) or have used a super-tree approach (Daubin et Gouy, 2001; Creevey et al., 2004; Beiko, Harlow et Ragan, 2005; Pisani, Cotton et McInerney, 2007). Our simulations suggest that this fear of HGTs is not justified when a large number of genes (100 or 200) are used. This number can be easily obtained using completely sequenced prokaryotic genomes (except for highly reduced genomes such as *Carsonella*). When HGTs are heterogeneously distributed across genes, an almost perfect accuracy of the super-matrix approach (98%) is obtained with 30 HGTs per gene on average (fig. 3.4). Although estimating the rate of HGTs is difficult, this value seems to be greater than what has been estimated thus far. For instance, less than 10 HGT events per gene were estimated over the proteomes of 144 species (Beiko et Hamilton, 2006). Another method that was focusing on the detection of HGTs within gene families, discovered no more than 10% of families showing evidence of transfer events. The authors inferred that only 1.1% of families having HGT between lineages of high taxonomic range, against 5.3% for

medium-level taxonomic range and 9.7% between closely related species (Choi et Kim, 2007).

Super-matrix seems to be slightly more accurate than the two super-tree methods tested when HGT rates are within a realistic range. The difference between super-trees (SDM and SuperTriplets) is much more important than between e.g. SDM and super-matrix. As a result, other super-tree algorithms should be tested, but no major surprises are expected since a similar result was obtained with 10 different super-tree methods when incongruence was generated by incomplete lineage sorting (Kupczok, Schmidt et von Haeseler, 2010). It has been argued that super-trees should be preferred because super-matrix ignores the different substitution rates and branch lengths across genes (Ren, Tanaka et Yang, 2009). But this branch length heterogeneity is simulated by the protocol of Galtier (2007), and the accuracy of the super-matrix approach is high despite the fact that we did not use models that allow different branch lengths across genes (Yang, 1996). Our results therefore suggest that the use of super-trees is likely not the best option for inferring the species tree, although this approach allows extracting conflicting signals, such as endosymbiotic gene transfers (Daubin et Gouy, 2001; Creevey et al., 2004; Beiko, Harlow et Ragan, 2005; Pisani, Cotton et McInerney, 2007).

More importantly, super-tree approaches seems to be more sensitive to tree reconstruction artifacts, especially long-branch attraction, than super-matrix (Philippe et al., 2005). Currently, artifacts due to model violations turn out to be the main limitation of phylogenomics in absence of HGTs, as clearly illustrated in the case of plants (Soltis et al., 2004; Stefanovic, Rice et Palmer, 2004; Rodriguez-Ezpeleta et al., 2007a) or animals (Lartillot, Brinkmann et Philippe, 2007; Nishihara, Okada et Hasegawa, 2007; Philippe et al., 2011a; Philippe et al., 2011b; Rota-Stabelli et al., 2011; Roure et Philippe, 2011). Since the phylogeny of prokaryotes is much deeper, artifacts are likely to play a more important role. Modeling the heterogeneity of the evolutionary process across sites (Lartillot et Philippe, 2004; Pagel et Meade, 2004) and/or over time (Galtier et Gouy, 1995; Foster, 2004; Blanquart et Lartillot, 2006; Blanquart et Lartillot, 2008) has lead to important improvements of the phylogenetic accuracy. However, these approaches apply more easily to super-matrix than to super-tree approaches, because

they require numerous positions to learn parameter values, in particular the global compositional trend. Albeit a joint estimation of independent tree topology with shared global parameters is possible, no implementation is yet available. Therefore, the super-matrix approach should be currently preferred over super-tree for inferring the prokaryotic tree of life using phylogenomics given its resilience to HGTs and the possibility to use complex models of sequence evolution.

Studying tree reconstruction artifacts using simulations is difficult, the introduced model violations are small with respect to the complex and large heterogeneity of real data. We use a different site-homogeneous model for simulation and for inference (JTT versus LG), and include large rate variation among genes through the parameter ρ'. Please note that this rate heterogeneity induces a model violation only for the super-matrix approach, since a separate model (Yang, 1996) was not used; this could have slightly decreased its accuracy. However, these model violations have a negligible effect as evidenced by the higher accuracy of phylogenies based on long single genes (fig. S3.3) and the very similar results obtained for different values of ρ' (fig. S3.4). To increase the effect of model violations, simulations were also performed with CAT+GTR, a complex site-heterogeneous model of sequence evolution (Lartillot et Philippe, 2004), which has usually a much better fit to large alignments than site-homogeneous models (Sperling, Peterson et Pisani, 2009; Philippe et al., 2011a; Rota-Stabelli et al., 2011). However, the results are similar to the original ones (fig. S3.9). The use of real data and of different models seems therefore necessary to study the effect of model violations more precisely.

As in Galtier (2007), the hypothesis that HGTs occur randomly was made for all our simulations. This simplifying assumption is certainly incorrect. It has been suggested that HGTs preferentially occur among closely related species (Gogarten, Doolittle et Lawrence, 2002). Not only this hypothesis does not seem to be supported (Puigbo, Wolf et Koonin, 2010), but phylogenetic accuracy would be less affected by these biased HGTs than by random HGTs. More problematic is the fact that HGTs are frequent in organisms thriving in the same habitat. This has been clearly established for environments at high temperature (Deckert et al., 1998; Nelson et al., 1999) and high

salinity (Mongodin et al., 2005). This bias was recently studied through simulation and was shown to limit the accuracy of the inference of the species tree (Beiko, Doolittle et Charlebois, 2008), however, only with a phenetic-like method, based on blast similarity scores (Clarke et al., 2002). Hence, testing more complex and probably more accurate super-tree and super-matrix methods remains to be done. However, species regularly change their habitats (except for a few extreme environments) and therefore habitat-driven HGTs could be relatively random in the long run, thereby making potentially the hypothesis of random HGTs a reasonable approximation. Further studies with more realistic models of HGTs are nevertheless needed.

In conclusion, our simulations show that, when numerous genes are considered, the use of genes with relatively high HGT rates has a limited impact on the inferred species tree, under reasonable assumptions about the evolutionary process. To make full use of the numerous complete genome sequences, we recommend inferring the prokaryotic phylogeny using hundreds of genes and a super-matrix with a complex model of sequence evolution. Other problems, due to missing data (i.e., gene loss) are possibly more important than previously estimated ((Lemmon et al., 2009) and unpublished results) and need also to be addressed. This approach should be efficient in inferring most of the trunk (the species tree) and will facilitate reconstituting the deviating part of the evolutionary history of single gene trees (the lianas), and thereby infer the HGT events that happened during the history of single gene families.

3.6 Supplementary material

Supplementary material concerning the different methods tested (fig. S3.1), the reference tree topology used (fig. S3.2 and see Supp. Materials) and how it has been obtained and supplementary information concerning results obtained for this paper are available.

3.7 Acknowledgments

We wish to thank Nicolas Lartillot for computational resources. We also thank the Réseau Québécois de Calcul de Haute Performance for computational resources.

H.P. was supported by the Canadian Research Chair Program, H.P. and H.B. were supported by NSERC and J.C.G was supported by CIHR and the Université de Montréal. We also want to thank Nicolas Galtier for helpful advices regarding the simulation procedure.

Supplementary material

An archaeal phylogeny was used as the reference tree of the simulation studies, it was inferred by RAxML (Stamatakis, 2006) under a LG+F+Γ4 model using a concatenation of 50 archaeal species with 20,052 amino acid positions. The dataset corresponds to the 47 biggest proteins of a larger data set of putative orthologous genes (table S2). To construct this larger data set, 79 complete genomes were downloaded from different public databases (NCBI: http://www.ncbi.nlm.nih.gov/sites/genome, JGI: http://www.jgi.doe.gov/genome-projects/, GOLD: http://www.genomesonline.org/ (Liolios et al., 2008)) and putative orthologous clusters were computed by OrthoMCL (Li, Stoeckert et Roos, 2003). This resulted in a total of 18,936 clusters of putative orthologous genes, but only 614 clusters with at least 50 species out of 79 were kept. Finally, we selected 50 species that represent the archaeal diversity in order to limit the computational costs.

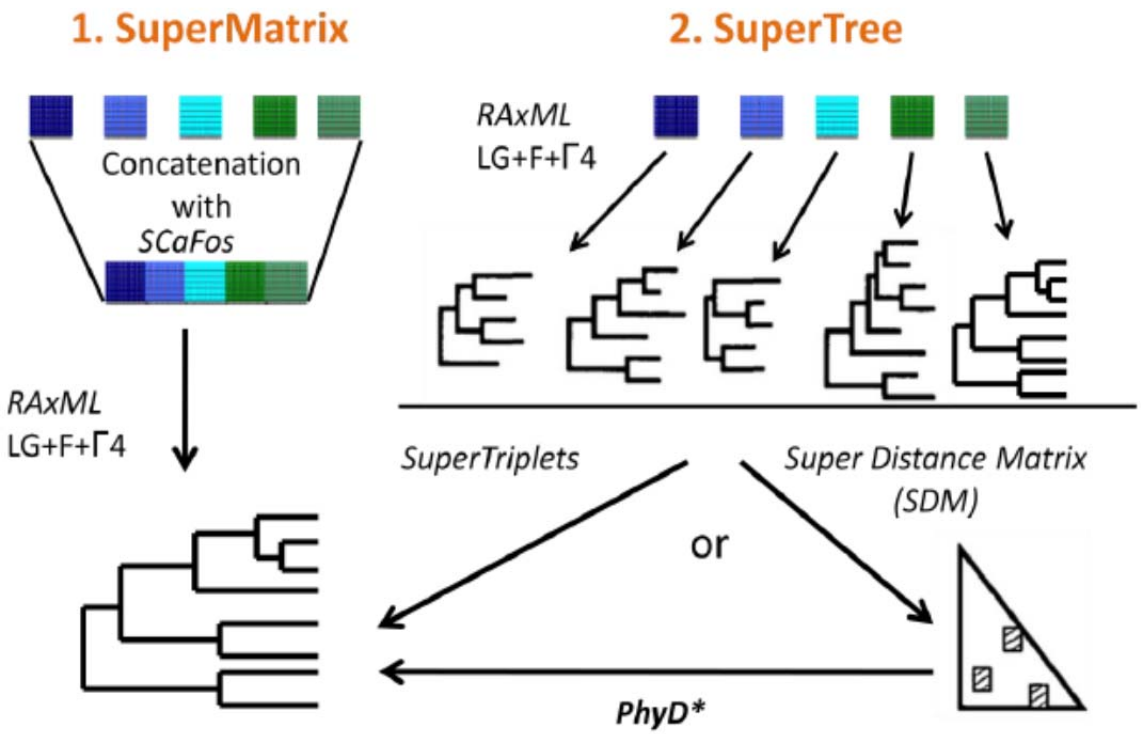


Figure S3.1 Flowchart of the two main phylogenomic approaches. 1- The super-matrix approach consists in the concatenation of all selected genes. The concatenation was performed with the program SCaFoS (Roure, Rodriguez-Ezpeleta et Philippe, 2007) and the ML tree inferences with RAxML (Stamatakis, 2006) under the LG (Le et Gascuel, 2008) model of sequence evolution with empirical amino acid frequencies and 4 discrete gamma categories. 2- The super-tree approaches consist of inferring individually all gene trees and then combine them in order to get a consensus tree/topology. The SuperTriplets (Ranwez, Criscuolo et Douzery, 2010) and the SDM (Criscuolo et al., 2006) methods were used.

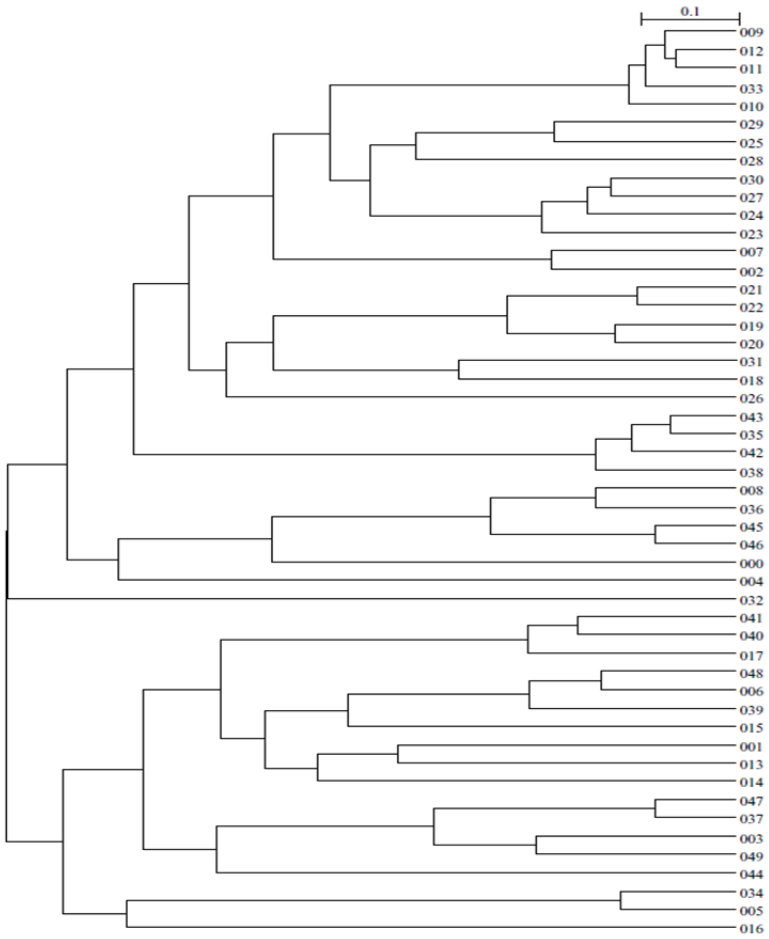


Figure S3.2 Reference tree used to perform HGT simulations (Galtier, 2007). The topology approximates the evolutionary history of Archaea. Note that the Newick format of the tree is available for reproduction of the experiment.

(((016:0.616051,(005:0.116061,034:0.116061):0.49999):0.0635434,((044:0.52395,((049:0.2014
3,003:0.20143):0.103211,(037:0.08086,047:0.08086):0.223781):0.21931):0.0741276,(((014:0.4
22636,(013:0.342019,001:0.342019):0.0806168):0.053011,(015:0.391715,(039:0.208623,(006:0
.1349,048:0.1349):0.0737235):0.183092):0.0839318):0.0440735,(017:0.209945,(040:0.159944,
041:0.159944):0.0500011):0.309775):0.0783578):0.0815165):0.0588618,(032:0.735411,((004:0
.62314,(000:0.468952,((046:0.0810049,045:0.0810049):0.16638,(036:0.141956,008:0.141956):
0.105429):0.221567):0.154187):0.0513705,((038:0.142177,(042:0.104457,(035:0.0667909,043:
0.0667909):0.0376659):0.03772):0.465721,((026:0.51454,((018:0.279678,031:0.279678):0.1877
53,((020:0.123326,019:0.123326):0.108618,(022:0.0997466,021:0.0997466):0.132198):0.23548
6):0.0471102):0.0379415,((002:0.187236,007:0.187236):0.27912,(((023:0.195973,(024:0.15041
4,(027:0.126332,030:0.126332):0.0240815):0.0455592):0.173559,(028:0.322008,(025:0.18348,
029:0.18348):0.138528):0.0475246):0.0402185,(010:0.108907,(033:0.090627,((011:0.0601338,
012:0.0601338):0.0117183,009:0.0718521):0.0187748):0.0182798):0.300844):0.056605):0.086
1261):0.0554161):0.066612):0.0609007):0.00304559);

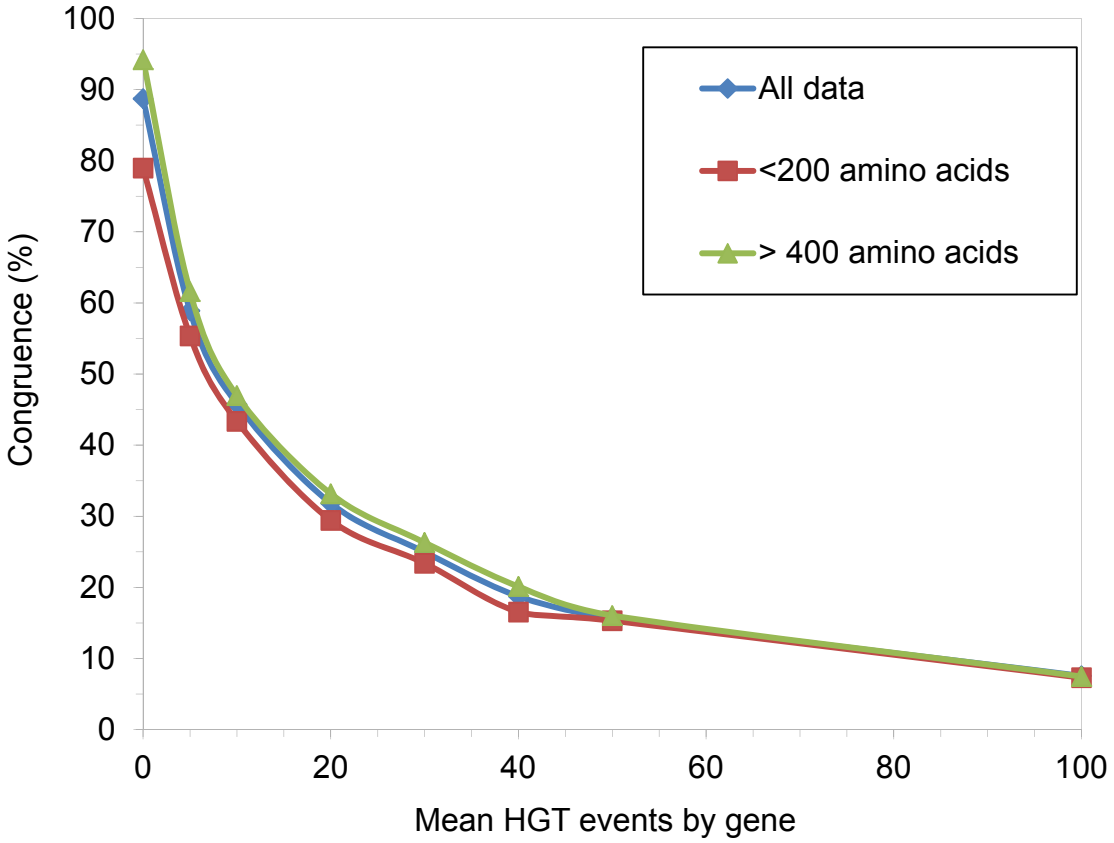


Figure S3.3 Comparison of the congruency of inferred gene trees compared to the reference tree, obtained for small and large proteins (compared to fig. 3.1) simulated for diverse levels of HGT events. The figure shows the results from the analysis of ten simulated genomes each with 100 genes with 20 genomic rate change (ρ) and 20 gene-specific rate change events (ρ'). Genomes of 50 species were simulated and each single gene has a randomly specified taxon sampling due to a random loss of taxa.

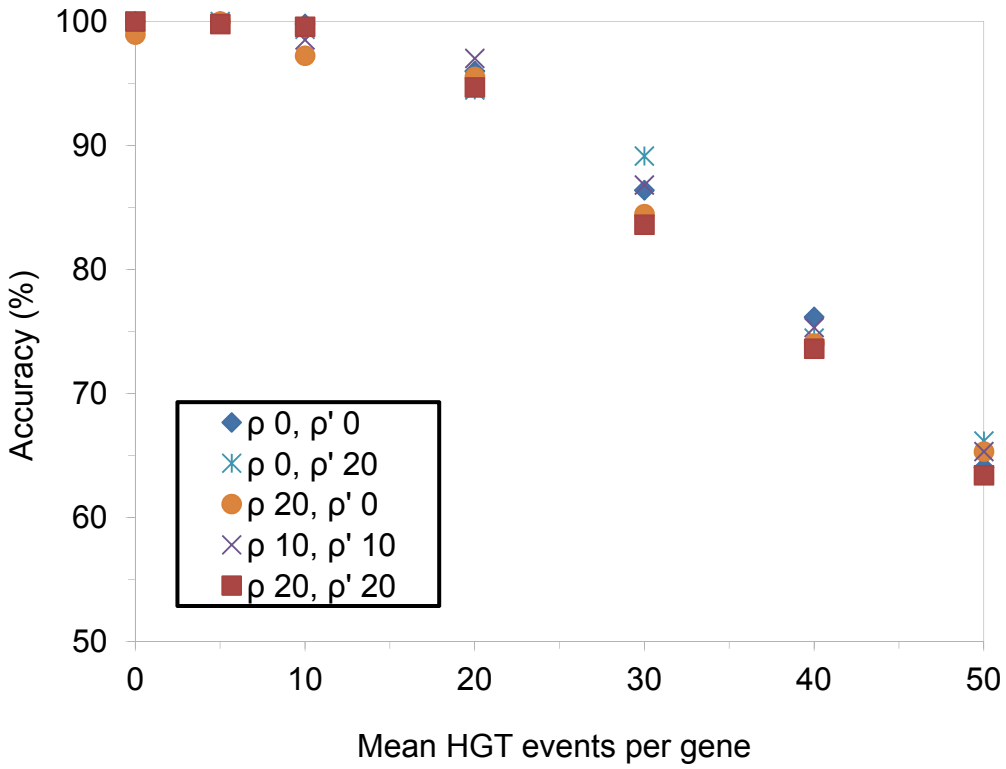


Figure S3.4 A comparison of the trees inferred from different simulated data sets containing various numbers of genomic (ρ) and gene specific (ρ') rate changes and inferred from a super-matrix approach. The different sets of trees were inferred using RAxML (Stamatakis, 2006) under a LG+Γ4 site-homogeneous model of sequence evolution. Each dot represents 10 replicates made for 100 genes combined with the inference approach.

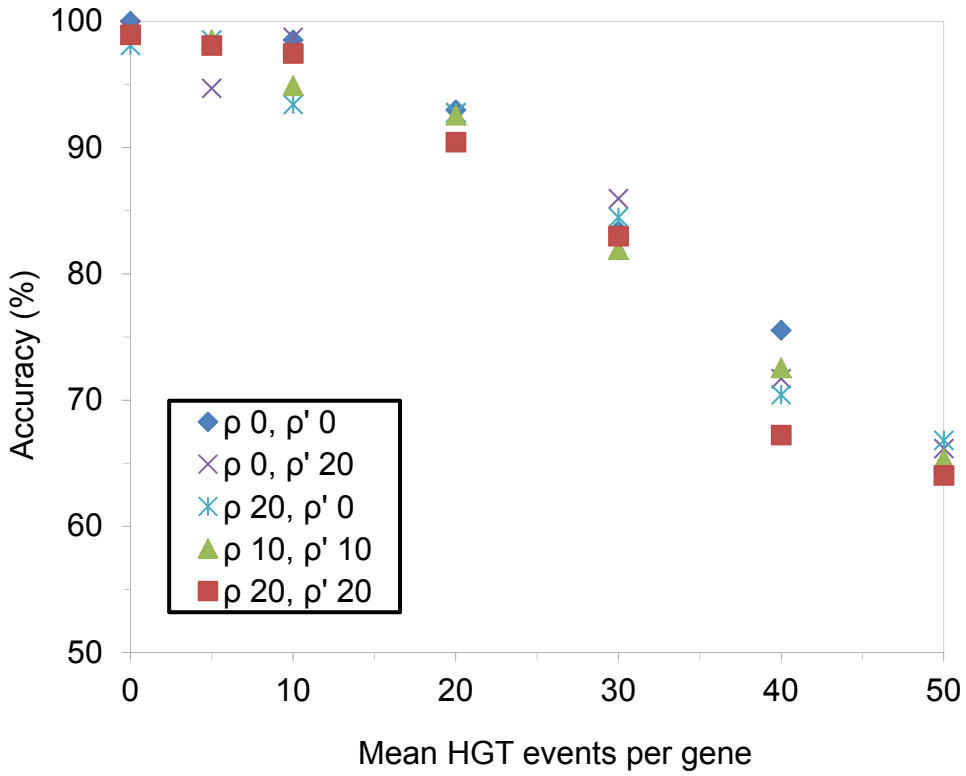


Figure S3.5 A comparison of the trees inferred from different simulated data sets containing a various number of genomic (ρ) and gene specific (ρ') rate changes and inferred by the SDM approach. The different sets of trees were inferred using RAxML (Stamatakis, 2006) under a LG+Γ4 site-homogeneous model of sequence evolution. Each dot represents 10 replicates made for 100 genes combined with the inference approach.

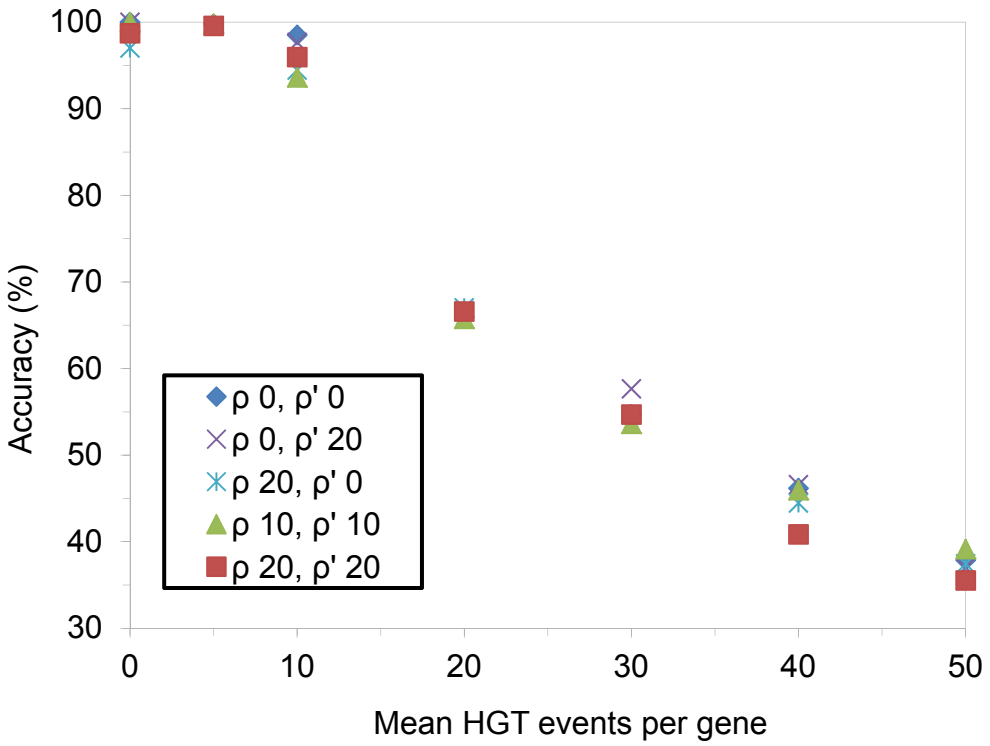


Figure S3.6 A comparison of the trees inferred by the SuperTriplets approach from simulated data sets containing different levels of genomic (ρ) and gene specific (ρ') rate changes. The different sets of trees were inferred using RAxML (Stamatakis, 2006) under a LG+Γ4 site-homogeneous model of sequence evolution. Each dot represents 10 replicates made for 100 genes combined with the inference approach.

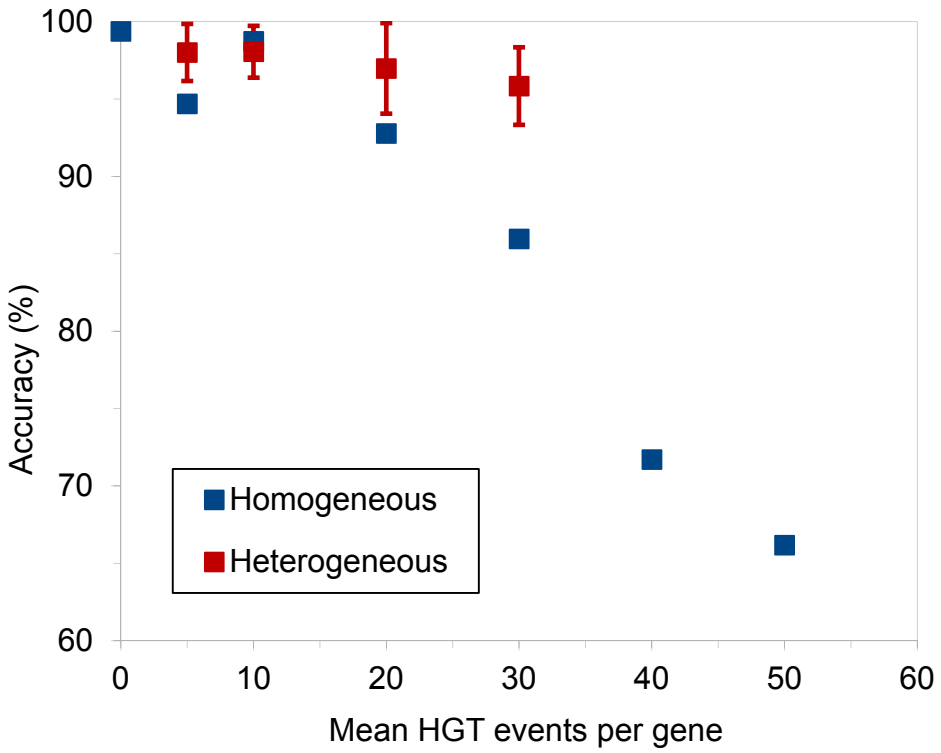


Figure S3.7 Comparison using the SDM approach of the effect of homo- and heterogeneous rates of HGTs per gene on the accuracy. The curves represent the accuracy of the estimated species trees in comparison to the reference tree that was used to do the simulations. The simulation parameters were ρ=0 and ρ'=20 based on 10 replicates per data-point with each dataset consisting of 100 genes. The combinations of different HGT levels used in these analyses are described in table 1.

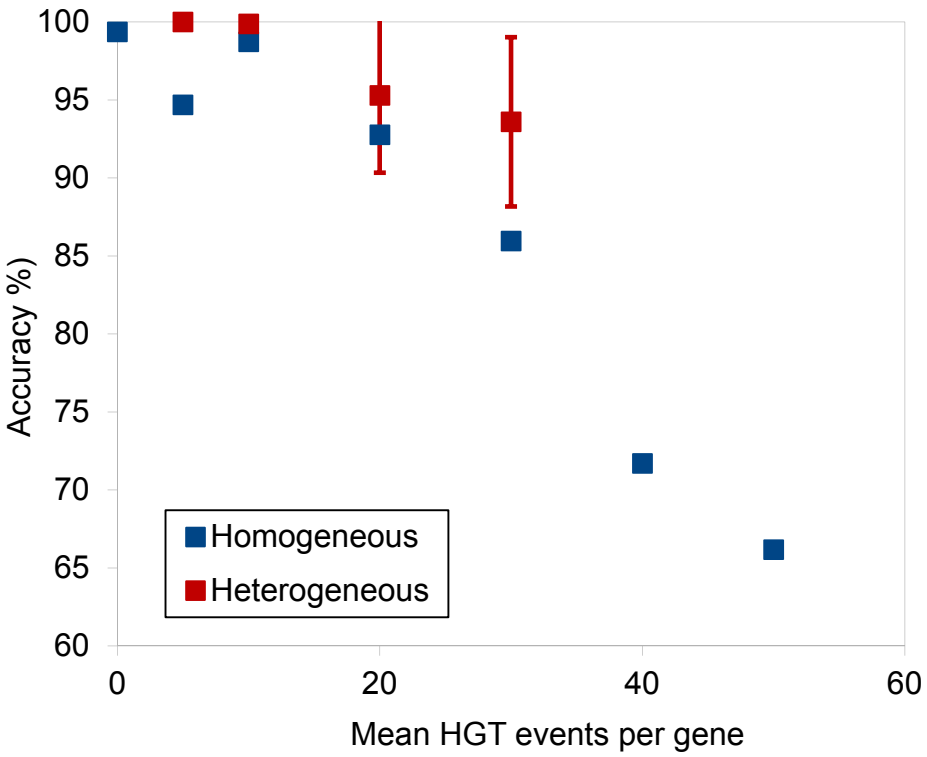


Figure S3.8 Comparison using the SuperTriplets approach of the effect of homo- and heterogeneous rates of HGTs per gene on the accuracy. The curves represent the accuracy of the estimated species trees in comparison to the reference tree that was used to do the simulations. The simulation parameters were ρ=0 and ρ'=20 based on 10 replicates per data-point with each dataset consisting of 100 genes. The combinations of different HGT levels used in these analyses are described in table 1.

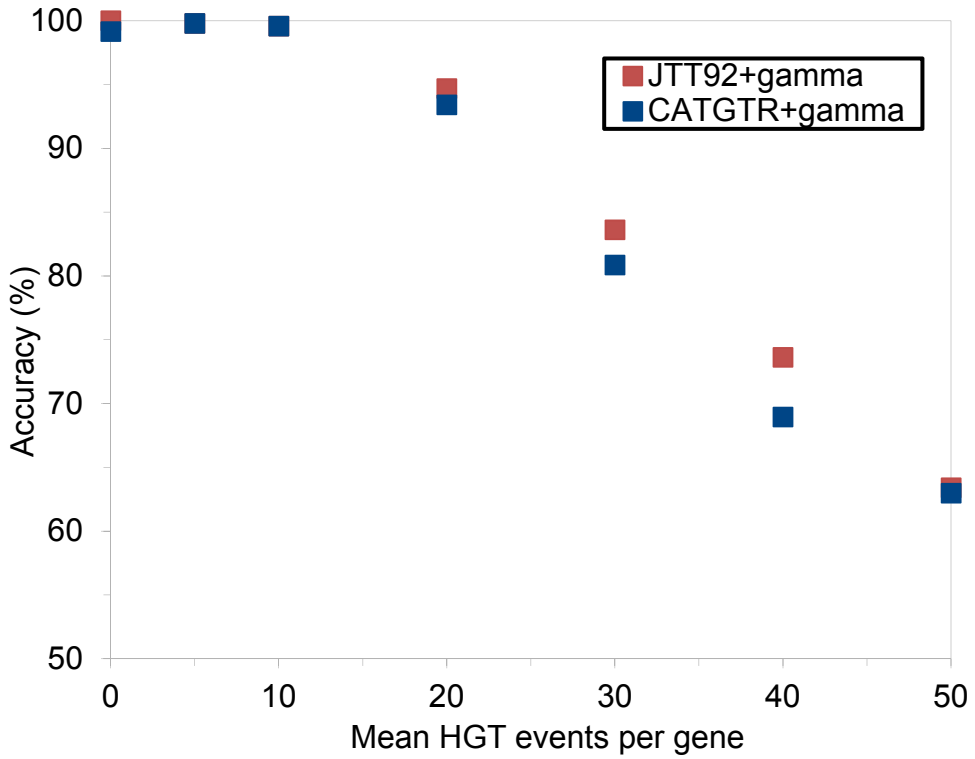


Figure S3.9 A comparison of the trees inferred from simulations using two different models of sequence evolution, JTT92+Γ and CATGTR+Γ. The two sets of trees were inferred using RAxML (Stamatakis, 2006) under a LG+Γ4 site-homogeneous model of sequence evolution. Each dot represents 10 replicates made for 100 genes combined with the inference approach.

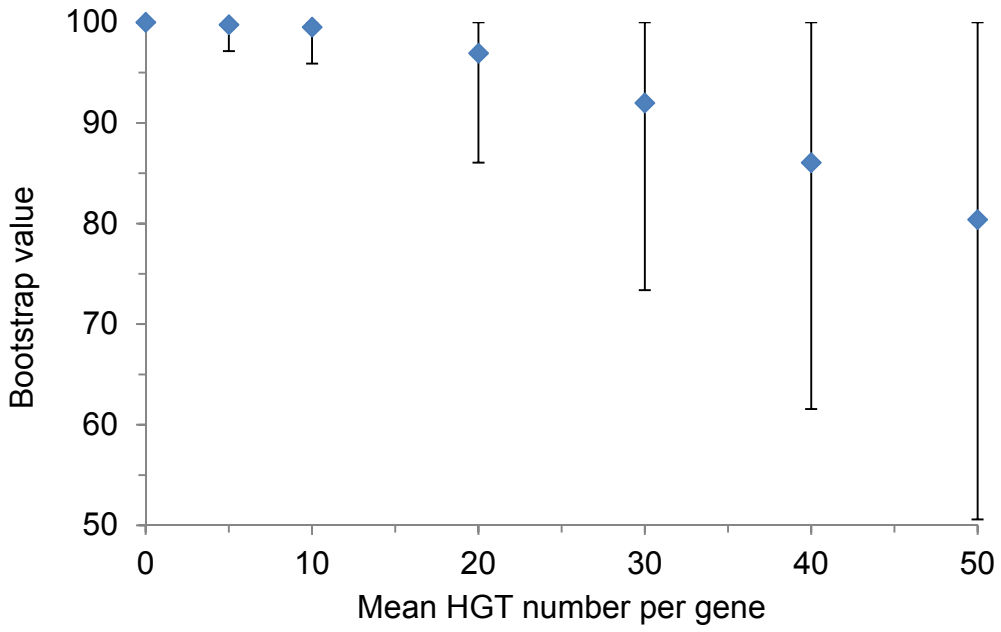


Figure S3.10 Bootstrap support values of the trees of figure 3.2 for the super-matrix approach. Each point represents 10 replicates containing 47 nodes.

Table S3.1 Comparison of parameters used to perform the HGT simulated data sets for our study and the one of N. Galtier (Galtier, 2007).

Parameter	Galtier (2007)	Our study
Number of species	40	50
Number of genes	20	5 to 200
Sequence length	100-500	100-500
Species Sampling Effort	50%-100%	66%-100%
d_k (tree diameter)	1-2-3	2
r (genome rate change)	0-8-16	0-10-20
r' (genome rate change)	0-8-16	0-10-20
t (HGT number per gene)	0-1-3-6-12	0-5-10-20-30-40-50-100
a_l (rate across lineages)	1	1
Substitution model	JTT	JTT
a_s (rate across sites)	0.5	0.5

Table S3.2 List of genes used to infer the reference tree. Those putative orthologous genes encode the 47 biggest proteins of a set of 18,936 putative orthologous genes and the corresponding super-matrix contains 20,052 amino acid positions.

Cluster ID	Function	Number of positions
c528	DNA polymerase, archaeal type II, large subunit	761
c134	DNA-directed RNA polymerase, subunit A'	670
c170	Isoleucyl-tRNA synthetase	618
c195	Translation elongation factor 2 (EF-2/EF-G)	616
c473	ATP dependent helicase, Lhr family	589
c457	Carbamoyl-phosphate synthase large subunit	572
c145	DNA-directed RNA polymerase, subunit B'	533
c345	Phosphoenolpyruvate synthase	528
c143	Alanyl-tRNA synthetase	522
c132	ATP synthase archaeal, A subunit	497
c236	Predicted ATPase, RNase L inhibitor (RLI) homolog	487
c581	Dihydroxyacid dehydratase	485
c127	n.d.	471
c600	n.d.	471
c188	Valyl-tRNA synthetase	467
c233	CTP synthase	442
c259	ATP synthase archaeal, B subunit	441
c680	Hydrogenase maturation protein HypF	431
c326	n.d.	417
c675	Pyruvate carboxylase, subunit A or Carbamoyl-phosphate synthase L chain, ATP-binding	417
c602	Chromosome segregation protein SMC	412
c633	Acetolactate synthase, large subunit	399
c368	Histone acetyltransferase, ELP3 family	396
c200	Translation initiation factor eaIF-5B	396
c380	ATP dependent helicase, Lhr family	395
c665	Heavy metal translocating P-type ATPase	394
c316	Adenosylhomocysteinase	385
c133	Translation elongation factor EF-1, subunit alpha	385
c128	Glutamyl-tRNA(Gln) amidotransferase subunit E	377
c343	Indolepyruvate ferredoxin oxidoreductase, alpha subunit	366
c162	Translation initiation factor IF-2 subunit gamma	358
c557	Type II secretion system protein E	358
c537	Fructose-1,6-bisphosphatase	352

c482	Ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent	348
c532	Pyridoxal-phosphate dependent TrpB-like enzyme	344
c232	ATPase, PilT family	343
c218	DEAD/DEAH box helicase domain protein	341
c175	Glycyl-tRNA synthetase	338
c231	Methionine adenosyltransferase or S-adenosylmethionine synthetase	335
c161	GTP-binding protein	333
c379	Phosphoribosylformylglycinamidine synthase subunit II	333
c530	Radical SAM domain protein	331
c194	Methionyl-tRNA synthetase	322
c121	Glutamine--fructose-6-phosphate transaminase	321
c629	Quinolinate phosphoribosyl transferase or Nicotinate-nucleotide pyrophosphorylase	320
c165	Glutamyl-tRNA synthetase	319
c371	Thiamine biosynthesis protein (thiC-1)	316

Conclusion

Les études qui ont été effectuées dans le cadre de ce mémoire ont permis d'estimer les effets des transferts horizontaux de gènes (THG) sur l'inférence de la phylogénie. Certaines hypothèses (Doolittle, 1999; Ge, Wang et Kim, 2005) stipulaient notamment que la notion d'arbre des espèces appliquée à l'histoire de la vie n'était plus envisageable à la lumière d'un niveau très élevé de THG. Le nombre de transmissions verticales a cependant été trouvé comme étant des milliers, voir des milliards de fois plus fréquentes que celui des transferts horizontaux (Philippe et Douady, 2003), nous laissant perplexe sur ces hypothèses. De plus, les méthodes de détection de THG n'étant pas encore complètement à point et donnant des résultats contradictoires (voir Chapitres 1.5.6 à 1.5.8), l'utilisation de plus de gènes n'est pas encore compromise. Les résultats obtenus par notre équipe dans cette direction concernant l'impact des THG ainsi que l'utilisation de modèles complexes d'évolution de séquences nous permettent de penser qu'il sera possible de continuer à améliorer la phylogénie des procaryotes en utilisant un plus grand nombre de gènes. Cela nous permet d'espérer qu'il sera possible de résoudre l'arbre de la vie lorsque les modèles et les ordinateurs seront beaucoup plus puissants. Il faut dire qu'une fraction non négligeable d'incongruences est très probablement liée à des problèmes dans les inférences phylogénétiques et non pas aux THG (Boc, Philippe et Makarenkov, 2010).

Avec les résultats obtenus, nous sommes parvenus à atteindre les objectifs que nous nous étions fixés (voir dans l'introduction) et les hypothèses ont pu être vérifiées. La réanalyse du jeu de données de protéines ribosomales (Chap. 2) à l'aide des modèles d'évolution de séquence site-hétérogène plus complexes (CAT+GTR+Γ_4, CAT+GTR+Dayhoff6+Γ_4) nous a permis d'identifier certains artéfacts de reconstruction phylogénomique. Quant aux simulations de THG (Chap. 3), ceux-ci auront permis d'étudier l'impact des THG sur l'inférence phylogénomique. Malgré le fait qu'un grand nombre de THG ait lieu pour chaque gène, la précision de la phylogénie ne s'en trouve pas trop affectée lorsqu'un grand nombre de gènes est utilisé, soit par une approche phylogénomique, c'est-à-dire qu'une perte de précision de 2% est observée lorsqu'il y a

par exemple 200 gènes utilisés et 20 THG en moyenne par gène pour un arbre contenant 50 espèces.

1 Vérifier la phylogénie des Archées

Quelques doutes persistaient concernant les phylogénies dernièrement publiées (notamment celle de Spang et al. (2010)) à cause de la position de quelques groupes, dont, en particulier, les Thaumarchaeota et les Korachaeota, qui sont considérés par beaucoup d'être des Crenarchaeota (DeLong, 1992; Fuhrman, McCallum et Davis, 1992; Schleper et al., 1997). Ils étaient considérés ainsi notamment à cause d'inférences effectuées sur l'ARN ribosomique 16S. Dans l'article de Spang et al. (2010), la position des Thaumarchaeota, par la grande branche qui est à leur base qui connecte les Archées avec le groupe externe (Eucaryotes), nous laissait croire qu'un artéfact de reconstruction puisse être à l'origine de cette position. L'idée d'effectuer une réanalyse de ce jeu de données nous est alors venue, dans l'espoir de pouvoir tester nos hypothèses concernant les causes du positionnement basal (et donc potentiellement erroné à cause de l'artéfact LBA) de ces groupes et de vérifier si la phylogénie actuelle des Archées était correcte. De plus, les positions et les longues branches menant aux Halobacteriales ainsi qu'à *Nanoarchaeum equitans* dans l'article de Spang et al. nous ont fait porter à croire à un artéfact de reconstruction, probablement une attraction de longues branches. Les différentes améliorations apportées dernièrement dans le domaine de la modélisation de l'évolution n'ont pas encore été testées dans chez les Archées. Il était alors nécessaire de vérifier si ces nouveaux modèles, représentant plus réalistement les tendances évolutives, pourraient obtenir les mêmes résultats que les modèles standards, considérant l'homogénéité substitutionnelle entre les sites.

1.1 Utilisation d'un modèle d'évolution de séquences complexe

La première topologie obtenue (Figure 2.1) confirme les différentes études effectuées avec des modèles complexes chez les animaux. La position des nœuds précédant les longues branches proches de la racine obtenue avec les modèles classiques

est probablement due à un artéfact de reconstruction phylogénétique, soit l'attraction des longues branches (ALB) plus précisément, causées par la présence d'un groupe externe éloigné en plus d'un groupe interne ayant une grande vitesse évolutive. Le modèle complexe utilisé, soit CAT+GTR+Γ_4, prenant en compte l'hétérogénéité par site (CAT) ainsi que l'hétérogénéité des patrons de substitution entre acides aminés (GTR), propose de placer les Thaumarchaeota comme groupe frère des Crenarchaeota. Il propose aussi de placer *Korarchaeum*, précédemment groupe frère des Euryarchaeota et des Crenarchaeota, ensemble avec les Crenarchaeota et les Thaumarchaeota. Ces résultats obtenus avec ces modèles complexes, corroborés par la validation croisée mentionnée dans le chapitre 2, viennent premièrement contredire certains récents résultats obtenus dans la littérature qui avaient été obtenus avec des modèles simples, tels que LG ou WAG. Ainsi, notre étude confirme l'importance pour l'étude de la phylogénomique d'utiliser de meilleurs modèles que les modèles sites homogènes standards étant donné la complexité des données génomiques. Les modèles complexes sont effectivement plus performants pour extraire le signal phylogénétique. Un grand nombre de modèles restent encore à tester. Nombre de ceux-ci, lorsque nous travaillons avec des jeux de données sensiblement imposants, nécessitent de très grandes ressources informatiques et sont encore trop lents afin d'obtenir des résultats satisfaisants dans des délais raisonnables. Les modèles utilisés lors de ce mémoire pourraient cependant être testés sur les jeux de données de Bactéries. Un protocole élaboré exposé en Annexe 1 pourrait d'ailleurs être utilisé afin de faire les différentes analyses phylogénomiques sur différents ensembles de données appartenant à différents phylums de Bactéries par exemple. Concernant les Archées, un meilleur échantillonnage taxonomique sera le meilleur moyen de pouvoir déterminer si ces regroupements sont supportés ou non.

1.2 Méthode de recodage des acides aminés

Une autre méthode visant à réduire les erreurs de reconstruction a également été utilisée. Suite à une analyse en composantes principales sur la composition en acides aminés de l'alignement, certaines espèces sont ressorties du lot, dont les Halobacteriales, *Methanopyrus kandleri* ainsi que *Nanoarchaeum equitans*. Ces points éloignés suggéraient la présence d'un biais compositionnel chez ces différentes espèces. Ce biais

compositionnel vient affecter la plupart des méthodes d'inférences en regroupant à tort les groupes ayant une composition similaire en acides aminés ou en G+C (Woese et al., 1991; Lockhart et al., 1994). Afin d'amoindrir ce biais, le recodage des acides aminés Dayhoff6 (Dayhoff, 1979) a été utilisé et a amené quelques changements topologiques. Le recodage utilisé permet d'éliminer certains biais compositionnels en regroupant les acides aminés en six classes fonctionnelles différentes, ne changeant pas les propriétés majeures des acides aminés pour différentes positions dans l'alignement protéique. Les Halobacteriales, qui sont précédés d'une très grande branche, sont désormais groupe frère des Methanomicrobiales et *Nanoarchaeum equitans* est désormais groupe frère des Thermococcales. Par contre, quelques incongruences sont apparues probablement dues à un manque de signal dans les données créé par le recodage. Cette perte de signal est causée par un changement dans le nombre d'états de caractères dû au recodage des acides aminés et à un nombre plutôt faible de positions (<7000). Les méthanogènes de type I (Methanobacteriales, Methanococcales et Methanopyrales) ont notamment été séparés en un groupe paraphylétique alors qu'ils sont probablement monophylétiques. La monophylie des Methanobacteriales et des Methanococcales est soutenue avec un support BP de 77% dans (Spang et al., 2010)(Figure 1.7) et avec un support de 1 en probabilité postérieure (PP) sur notre arbre CAT+GTR+Γ4 (Figure 2.1), alors que la paraphylie est supportée à plus de 0.9 PP avec la technique de recodage (Figure 2.2). Les résultats obtenus à l'aide de ces modèles plus complexes laissent planer une incertitude par rapport aux topologies obtenues dans les articles parus récemment dans la littérature (Brochier-Armanet et al., 2008; Spang et al., 2010). Les modèles complexes expliquant mieux les données que le modèle utilisé par Spang et al. (valeur de "fit" : -1499±88 sur 10 réplicats de la validation croisée), nous croyons cependant que la topologie obtenue à l'aide de ces modèles est plus proche de la réalité.

Avec les changements topologiques observés et les caractéristiques compositionnelles des différents génomes, il semble évident que l'utilisation exclusive d'un modèle simple n'est plus suffisante afin de travailler à un niveau taxonomique élevé. Beaucoup de travail est également nécessaire afin de rendre plus rapides les nouveaux modèles complexes, dits site-hétérogènes, puisque des analyses de grands jeux de données peut prendre un temps très élevé à cause du nombre très élevé de paramètres

à calculer. Comme ils sont conçus pour le moment, l'augmentation du nombre de sites complexifie la détermination du nombre de patrons de substitution à générer, pour le modèle CAT entre autres, et une augmentation du nombre d'espèces fait bien sûr augmenter le nombre de paramètres à calculer ainsi que toutes les longueurs de branches et la topologie elle-même qui s'en trouve affectée par sa complexité. Une augmentation du nombre de sites utilisés ne sera par contre pas négative lorsqu'il sera question d'utiliser un système de recodage d'acides aminés, puisqu'il y a beaucoup moins d'états de caractères à prendre en compte. Il serait également possible de tester différents autres types de recodage personnalisés (Susko et Roger, 2007).

2 Impact des transferts horizontaux de gènes

Les manipulations effectuées dans ce mémoire avaient pour but de vérifier l'impact qu'auraient différents taux de THG sur les approches phylogénomiques, puisque ceci n'avait pas encore été testé dans la littérature. Deux approches de super-arbre et une approche super-matrice, se basant sur un modèle site-homogène ont été testées sur des simulations. Même avec ce modèle (LG), qui n'est pas le plus puissant, la précision avec laquelle la topologie de référence était retrouvée s'est améliorée avec l'augmentation du nombre de positions. De plus, l'approche super-matrice, qui est majoritairement utilisée en phylogénomique, démontre une bonne robustesse au problème des THG, en étant capable d’extraire une grande partie du signal lié à la transmission verticale, et ce même si le taux de THG est exagéré par rapport à ce qui est observé sur des données réelles. D'autres approches pourraient également être mises en œuvre afin de vérifier l'impact des THG. Un outil a notamment été élaboré au cours de ma Maîtrise (voir Annexe 2). Cet outil permet de détecter les THG potentiels (ou les séquences évoluant très vite ou très lentement) et les retire des alignements. De cette façon, le signal en faveur des THG (et des artéfacts de reconstruction) est retiré, ce qui permettrait d'augmenter la robustesse du signal vertical des phylogénies inférées. Par contre, ceci entraîne une augmentation des données manquantes, un point qui a été plus approfondi par Béatrice Roure, membre de notre laboratoire. Celle-ci démontre que l'augmentation des données manquantes peut affecter la fiabilité des phylogénies inférées de façon assez prononcée plus le taux de

celles-ci augmente et que celles-ci sont réparties de façon non-aléatoire (prochainement soumis).

3 Perspectives

Les résultats obtenus dans ce mémoire laissent prévoir une potentialité certaine de continuer à travailler sur la phylogénie des procaryotes, représentés par les Bactéries et les Archées. Les modèles d'évolution de séquences complexes s'étant montrés aussi utiles chez ceux-ci que chez les Eucaryotes, nous croyons que d'autres modèles d'évolution de séquences conçus spécialement pour le cas des THG (Novozhilov, Karev et Koonin, 2005) seraient d'une grande utilité afin d'obtenir encore plus de robustesse pour cette phylogénie. Les résultats obtenus dans ce mémoire sur la phylogénie des Archées mettent surtout l'emphase sur le besoin d'utiliser de tels modèles complexes de nos jours afin d'obtenir une meilleure résolution et ceci combiné avec la disponibilité de beaucoup plus de génomes complets chez ces organismes. Bien qu'il n'ait pas été étudié dans ce mémoire, le domaine des Bactéries, pour lesquelles le nombre de génomes complets connus est beaucoup plus grand que les Archées, serait intéressant à étudier avec ces nouveaux modèles. La structure phylogénétique de ce domaine pourrait ainsi être révisée et améliorée. Des analyses de jeux de données, se basant sur les protéines ribosomales de tous les nouveaux génomes complètement séquencés ou d'un grand nombre de protéines de natures différentes, pourraient ainsi être effectuées à l'aide de ces nouveaux modèles.

Suivant le raisonnement de certains auteurs, l'utilisation d'un faible nombre de gènes peut cependant être douteuse si on considère que le génome complet de ces espèces est disponible et que ce nombre de gènes ne représente qu'un pour cent du génome moyen de procaryote (Dagan et Martin, 2006). Les résultats obtenus concernant les simulations de THG viennent donner de l'espoir à la phylogénomique dans son effort de reconstruire l'arbre évolutif de ces groupes ainsi que l'arbre de la vie. Le fait qu'un nombre moyen très élevé de transferts par gène (20 THG) affecte peu la précision obtenue pour l'arbre des espèces inféré, par rapport à l'arbre de référence, laisse supposer qu'un grand nombre de gènes (>100) pourraient dorénavant être utilisés afin d'obtenir la phylogénie d'un

groupe de procaryotes potentiellement affecté par des THG. Cette phylogénie représenterait de ce fait une bonne fraction du génome des espèces à la place d'un pour cent. Bien sûr, il s'agirait de sélectionner des gènes quasi universels en plus de gènes contenant une majorité des espèces étudiées. Notre étude concernant l'impact des THG étant basé sur un modèle assez simple, d'autres simulations comportant différents types de biais pourraient être effectuées afin d'améliorer le réalisme dans les simulations. Un scénario de THG uniformes entre les espèces et dans le temps est effectivement peu probable. Il s'agirait de vérifier l'impact des THG sur les approches phylogénomiques à l'aide d'une technique comme Beiko et al. (2008) qui a simulé des THG biaisés entre espèces proches, entre génomes ayant une composition similaire et entre espèces vivant dans des habitats semblables. Ceci recréerait ainsi le fait exposé que des THG sont plus susceptibles lorsqu'une espèce change d'habitat. Pour aller encore plus loin, nous pourrions effectuer des simulations en spécifiant explicitement des espèces étant connues pour s'échanger très souvent des gènes et même y inclure des espèces provenant des deux autres domaines, donc des espèces très éloignées. Les THG pourraient également être effectuées en ne considérant pas leur distribution uniforme au cours du temps, mais en fixant certaines périodes cibles, soient différentes périodes de diversification des espèces, etc. De cette façon, l'ensemble des conditions dans lesquelles les THG pourraient survenir serait inspecté.

Les nouvelles analyses phylogénomiques effectuées chez les procaryotes avec un grand nombre de gènes pourraient venir aider à établir une meilleure vision et une meilleure compréhension du phénomène des transferts horizontaux de gènes. Cela pourrait représenter une bonne façon d'améliorer les outils de détection de THG existants ou à venir. Une phylogénie de référence basée sur une grande quantité de gènes (>100) et étant très robuste serait selon notre hypothèse une meilleure approximation de l'histoire évolutive du génome et serait de ce fait un arbre de référence beaucoup plus certain à utiliser pour les approches phylogénétiques. Dans le cadre de cette maitrise, un programme de détection de THG non-phylogénétique a été conçu, se basant sur la comparaison de matrices de distances. Celui-ci requiert par contre encore quelques ajustements et améliorations afin d'être complètement opérationnel. Sa description est disponible dans l'annexe 1 et un protocole détaillé est disponible dans l'annexe 2.

L'arbre des espèces inféré, représentant l'histoire évolutive globale des génomes, représenterait également un atout majeur afin de rebâtir l'histoire de certains gènes. Obtenir plus d'informations sur ceux-ci serait pertinent afin de mieux comprendre une partie de l'histoire de notre planète liée à la présence de la vie et d'avoir une meilleure compréhension de l'évolution de la vie elle-même. La datation de l'apparition de certains traits phénotypiques en se basant sur des arbres phylogénomiques tels que la méthanogénèse ou encore le moment précis des différents événements d'endosymbioses permettrait de mieux cerner les différents épisodes de changements climatiques et environnementaux étant survenus sur la planète.

La monophylie des Archées et une meilleure définition du dernier ancêtre commun (LUCA) représentent également des sujets qui pourraient être avancés considérant l'existence de l'arbre de la vie. L'utilisation de plus de gènes, même si ceux-ci contiennent des THG, et de modèles complexes prenant en considération beaucoup de concepts évolutifs seront alors de mise afin d'obtenir un arbre avec assez de résolution pour pouvoir obtenir des réponses à ces questions fondamentales.

Bibliographie

Abby, S. S., E. Tannier, M. Gouy and V. Daubin. 2010. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinformatics. 11: 324.

Akerborg, O., B. Sennblad, L. Arvestad and J. Lagergren. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci U S A. 106(14): 5714-5719.

Allen, B. L. and M. A. Steel. 2001. Subtree transfer operations and their induced metrics on evolutionary trees. Ann. Combinatorics. 5: 1-15.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol. 215(3): 403-410.

Aravalli, R. N., Q. She and R. A. Garrett. 1998. Archaea and the new age of microorganisms. Trends in Ecology & Evolution. 13(5): 190-194.

Baker, B. J., L. R. Comolli, G. J. Dick et al. 2010. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A. 107(19): 8806-8811.

Bapteste, E., H. Brinkmann, J. A. Lee et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. USA. 99(3): 1414-1419.

Bapteste, E., Y. Boucher, J. Leigh and W. F. Doolittle. 2004. Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol. 12(9): 406-411.

Bapteste, E., E. Susko, J. Leigh, D. MacLeod, R. L. Charlebois and W. F. Doolittle. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evol Biol. 5(1): 33.

Bapteste, E., M. A. O'Malley, R. G. Beiko et al. 2009. Prokaryotic evolution and the tree of life are two different things. Biol Direct. 4: 34.

Barns, S. M., C. F. Delwiche, J. D. Palmer and N. R. Pace. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci U S A. 93(17): 9188-9193.

Baum, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon. 41: 3–10.

Baurain, D. and H. Philippe. 2010. Current Approaches to Phylogenomic Reconstruction.In: G. Caetano-Anollés. Evolutionary Genomics and Systems Biology. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.**:** 17-41.

Beiko, R. G., T. J. Harlow and M. A. Ragan. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A. 102(40): 14332-14337.

Beiko, R. G. and N. Hamilton. 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evol Biol. 6: 15.

Beiko, R. G., W. F. Doolittle and R. L. Charlebois. 2008. The impact of reticulate evolution on genome phylogeny. Syst Biol. 57(6): 844-856.

Berg, O. G. and C. G. Kurland. 2002. Evolution of microbial genomes: sequence acquisition and loss. Mol Biol Evol. 19(12): 2265-2276.

Blanchette, M., G. Bourque and D. Sankoff. 1997. Breakpoint phylogenies. Eighth Genome Informatics Conference (GIW 1997), Universal Academy Press.

Blanchette, M., T. Kunisawa and D. Sankoff. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. J. Mol. Evol. 49(2): 193-203.

Blanquart, S. and N. Lartillot. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Mol Biol Evol. 23(11): 2058-2071.

Blanquart, S. and N. Lartillot. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol. 25(5): 842-858.

Bloomquist, E. W. and M. A. Suchard. 2010. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. Syst Biol. 59(1): 27-41.

Boc, A., H. Philippe and V. Makarenkov. 2010. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. Syst Biol. 59(2): 195-211.

Bordewich, M., C. Semple and J. Talbot. 2004. Counting consistent phylogenetic trees is #P-complete. Advances in Applied Mathematics. 33(2): 416-430.

Boucher, Y. and E. Bapteste. 2009. Revisiting the concept of lineage in prokaryotes: a phylogenetic perspective. Bioessays. 31(5): 526-536.

Boussau, B., S. Blanquart, A. Necsulea, N. Lartillot and M. Gouy. 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature. 456(7224): 942-945.

Brinkmann, H. and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16(6): 817-825.

Brinkmann, H., M. Giezen, Y. Zhou, G. P. Raucourt and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol. 54(5): 743-757.

Brochier-Armanet, C., B. Boussau, S. Gribaldo and P. Forterre. 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol. 6(3): 245-252.

Brochier, C., E. Bapteste, D. Moreira and H. Philippe. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends Genet. 18(1): 1-5.

Brochier, C., P. Forterre and S. Gribaldo. 2004. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. Genome Biol. 5(3): R17.

Brochier, C., P. Forterre and S. Gribaldo. 2005. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evol Biol. 5(1): 36.

Brochier, C., S. Gribaldo, Y. Zivanovic, F. Confalonieri and P. Forterre. 2005. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? Genome Biol. 6(5): R42.

Buchanan-Wollaston, V., J. E. Passiatore and F. Cannon. 1987. The mob and oriT mobilization functions of a bacterial plasmid promote its transfer to plants. Nature. 328(6126): 172-175.

Buckley, D. H., J. R. Graber and T. M. Schmidt. 1998. Phylogenetic analysis of nonthermophilic members of the kingdom crenarchaeota and their diversity and abundance in soils. Appl Environ Microbiol. 64(11): 4333-4339.

Burggraf, S., K. O. Stetter, P. Rouviere and C. R. Woese. 1991. *Methanopyrus kandleri*: an archaeal methanogen unrelated to all other known methanogens. Syst Appl Microbiol. 14: 346-351.

Burki, F., A. Kudryavtsev, M. V. Matz, G. V. Aglyamova, S. Bulman, M. Fiers, P. J. Keeling and J. Pawlowski. 2010. Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. BMC Evol Biol. 10: 377.

Campbell, A., J. Mrazek and S. Karlin. 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. USA. 96(16): 9184-9189.

Cavalier-Smith, T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int J Syst Evol Microbiol. 52(Pt 2): 297-354.

Cavicchioli, R. 2006. Cold-adapted archaea. Nat Rev Microbiol. 4(5): 331-343.

Charlebois, R. L., R. G. Beiko and M. A. Ragan. 2003. Microbial phylogenomics: Branching out. Nature. 421(6920): 217.

Choi, I. G. and S. H. Kim. 2007. Global extent of horizontal gene transfer. Proc Natl Acad Sci U S A. 104(11): 4489-4494.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311(5765): 1283-1287.

Clarke, G. D., R. G. Beiko, M. A. Ragan and R. L. Charlebois. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J Bacteriol. 184(8): 2072-2080.

Cohen, O. and T. Pupko. 2010. Inference and characterization of horizontally transferred gene families using stochastic mapping. Mol Biol Evol. 27(3): 703-713.

Cohen, O., U. Gophna and T. Pupko. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 28(4): 1481-1489.

Cook, R. 1979. Influential Observations in Linear Regression. Journal of the American Statistical Association(74): 169-174.

Cox, C. J., P. G. Foster, R. P. Hirt, S. R. Harris and T. M. Embley. 2008. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A. 105(51): 20356-20361.

Creevey, C. J., D. A. Fitzpatrick, G. K. Philip, R. J. Kinsella, M. J. O'Connell, M. M. Pentony, S. A. Travers, M. Wilkinson and J. O. McInerney. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc Biol Sci. 271(1557): 2551-2558.

Criscuolo, A., V. Berry, E. J. Douzery and O. Gascuel. 2006. SDM: a fast distance-based approach for (super) tree building in phylogenomics. Syst Biol. 55(5): 740-755.

Criscuolo, A. and O. Gascuel. 2008. Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics. 9: 166.

Csuros, M. 2006. On the estimation of intron evolution. PLoS Comput Biol. 2(7): e84; author reply e83.

Csuros, M. and I. Miklos. 2009. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol. 26(9): 2087-2095.

Dagan, T. and W. Martin. 2006. The tree of one percent. Genome Biol. 7(10): 118.

Dagan, T., Y. Artzy-Randrup and W. Martin (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution, National Academy of Sciences.

Darwin, C. 1859. The origin of species by means of natural selection. London: Murray.

Daubin, V. and M. Gouy. 2001. Bacterial molecular phylogeny using supertree approach. Genome Inform Ser Workshop Genome Inform. 12: 155-164.

Daubin, V., M. Gouy and G. Perriere. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. 12(7): 1080-1090.

Daubin, V., E. Lerat and G. Perriere. 2003. The source of laterally transferred genes in bacterial genomes. Genome Biol. 4(9): R57.

Daubin, V. and G. Perriere. 2003. G+C3 structuring along the genome: a common feature in prokaryotes. Mol Biol Evol. 20(4): 471-483.

Davies, J. and D. Davies. 2010. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev. 74(3): 417-433.

Dayhoff, M. O., R. V. Eck and C. M. Park. 1972. A model of evolutionary change in proteins.In: M. O. Dayhoff. Atlas of protein sequence and structure. Washington, DC: National Biomedical Research Fundation. **5:** 89-99.

Dayhoff, M. O. 1979. Atlas of Protein Sequence and Structure.In. Washington, D.C.: Supplement 3, 1978. National Biomedical Research Foundation. **5:** 345-352.

Dayrat, B. 2003. The roots of phylogeny: how did Haeckel build his trees? Syst Biol. 52(4): 515-527.

Deckert, G., P. V. Warren, T. Gaasterland et al. 1998. The complete genome of the hyperthermophilic bacterium Aquifex aeolicus. Nature. 392(6674): 353-358.

DeLong, E. F. 1992. Archaea in coastal marine environments. Proc Natl Acad Sci U S A. 89(12): 5685-5689.

Delport, W., K. Scheffler and C. Seoighe. 2009. Models of coding sequence evolution. Brief Bioinform. 10(1): 97-109.

Delsuc, F., M. J. Phillips and D. Penny. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?". Science. 301(5639): 1482; author reply 1482.

Delsuc, F., H. Brinkmann and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6(5): 361-375.

Doolittle, R. F. and J. Handy. 1998. Evolutionary anomalies among the aminoacyl-tRNA synthetases. Curr Opin Genet Dev. 8(6): 630-636.

Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. Science. 284(5423): 2124-2129.

Doolittle, W. F., Y. Boucher, C. L. Nesbo, C. J. Douady, J. O. Andersson and A. J. Roger. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos Trans R Soc Lond B Biol Sci. 358(1429): 39-57; discussion 57-38.

Doolittle, W. F. and E. Bapteste. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A. 104(7): 2043-2049.

Dopazo, H. and J. Dopazo. 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. 6(5): R41.

Douzery, E. J. P., S. Blanquart, A. Criscuolo, F. Delsuc, C. Douady, N. Lartillot, H. Philippe and V. Ranwez. 2010. Phylogénie moléculaire.In: F. Thomas, T. Lefevre and M. Raymond. Biologie évolutive. Bruxelles: De Boeck**:** 183-243.

Dunn, C. W., A. Hejnol, D. Q. Matus et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452(7188): 745-749.
Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5: 113.
Edwards, S. V., B. Fertil, A. Giron and P. J. Deschavanne. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. Syst. Biol. 51(4): 599-613.
Elkins, J. G., M. Podar, D. E. Graham et al. 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. Proc Natl Acad Sci U S A. 105(23): 8102-8107.
Faguy, D. M. and W. F. Doolittle. 1999. Lessons from the *Aeropyrum pernix* genome. Curr Biol. 9(23): R883-886.
Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27: 401-410.
Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17(6): 368-376.
Felsenstein, J. and G. A. Churchill. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13(1): 93-104.
Field, K. G., G. J. Olsen, D. J. Lane, S. J. Giovannoni, M. T. Ghiselin, E. C. Raff, N. R. Pace and R. A. Raff. 1988. Molecular phylogeny of the animal kingdom. Science. 239(4841 Pt 1): 748-753.
Fitch, W. M. and E. Margoliash. 1967. Construction of phylogenetic trees. Science. 155(760): 279-284.
Fitch, W. M. and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet. 4(5): 579-593.
Fitz-Gibbon, S. T. and C. H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. 27(21): 4218-4222.
Forterre, P. and H. Philippe. 1999. The last universal common ancestor (LUCA), simple or complex? Biol Bull. 196(3): 373-375; discussion 375-377.
Forterre, P., C. Bouthier De La Tour, H. Philippe and M. Duguet. 2000. Reverse gyrase from hyperthermophiles: probable transfer of a thermoadaptation trait from archaea to bacteria. Trends Genet. 16(4): 152-154.
Forterre, P., C. Brochier and H. Philippe. 2002. Evolution of the Archaea. Theor Popul Biol. 61(4): 409-422.
Foster, P. G. 2004. Modeling compositional heterogeneity. Syst Biol. 53(3): 485-495.
Foster, P. G., C. J. Cox and T. M. Embley. 2009. The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods. Philos Trans R Soc Lond B Biol Sci. 364(1527): 2197-2207.
Friend, T. 2007. The Third Domain: The Untold Story of Archaea and the Future of Biotechnology. Washington D.C.: Joseph Henry Press.
Frost, L. S., R. Leplae, A. O. Summers and A. Toussaint. 2005. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 3(9): 722-732.
Fuhrman, J. A., K. McCallum and A. A. Davis. 1992. Novel major archaebacterial group from marine plankton. Nature. 356(6365): 148-149.
Galtier, N. and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA. 92(24): 11317-11321.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol. 18(5): 866-873.
Galtier, N. 2007. A model of horizontal gene transfer and the bacterial phylogeny problem. Syst Biol. 56(4): 633-642.
Galtier, N. and V. Daubin. 2008. Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc Lond B Biol Sci. 363(1512): 4023-4029.
Ge, F., L. S. Wang and J. Kim. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3(10): e316.
Germot, A. and H. Philippe. 1999. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. J Eukaryot Microbiol. 46(2): 116-124.
Gogarten, J. P., H. Kibak, P. Dittrich et al. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA. 86(17): 6661-6665.
Gogarten, J. P., W. F. Doolittle and J. G. Lawrence. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol. 19(12): 2226-2238.
Gogarten, J. P. and J. P. Townsend. 2005. Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol. 3(9): 679-687.
Goldman, N., J. L. Thorne and D. T. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 149: 445-458.
Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl and F. H. Hellwig. 2003. Analysis of the Amborella trichopoda chloroplast genome sequence suggests that amborella is not a basal angiosperm. Mol Biol Evol. 20(9): 1499-1505.
Gribaldo, S. and C. Brochier-Armanet. 2006. The origin and evolution of Archaea: a state of the art. Philos Trans R Soc Lond B Biol Sci. 361(1470): 1007-1022.
Griffith, F. 1928. The Significance of Pneumococcal Types. J Hyg (Lond). 27(2): 113-159.
Gruenheit, N., P. J. Lockhart, M. Steel and W. Martin. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. Mol Biol Evol. 25(7): 1512-1520.
Gu, X. and H. Zhang. 2004. Genome phylogenetic analysis based on extended gene contents. Mol Biol Evol. 21(7): 1401-1408.
Haeckel, E. 1866. Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie. Berlin: Georg Reimer.
Hahn, M. W., T. De Bie, J. E. Stajich, C. Nguyen and N. Cristianini. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 15(8): 1153-1160.
Halpern, A. L. and W. J. Bruno. 1998. Evolutionary distances for protein-coding sequences: modeling site- specific residue frequencies. Mol. Biol. Evol. 15(7): 910-917.
Hao, W. and G. B. Golding. 2006. The fate of laterally transferred genes: life in the fast lane to adaptation or death. Genome Res. 16(5): 636-643.
Hao, W. and G. B. Golding. 2010. Inferring bacterial genome flux while considering truncated genes. Genetics. 186(1): 411-426.
Hasegawa, M., H. Kishino and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22(2): 160-174.

Hein, J., T. Jiang, L. Wang and K. Zhang. 1996. On the complexity of comparing evolutionary trees. Discrete Applied Mathematics. 71(1-3): 153-169.
Heinemann, J. A. and G. F. Sprague, Jr. 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. Nature. 340(6230): 205-209.
Hejnol, A., M. Obst, A. Stamatakis et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci.
Hendy, M. D. and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38: 297-309.
Hennig, W. 1966. Phylogenetic systematics. Urbana: University of Illinois Press.
Hilario, E. and J. P. Gogarten. 1993. Horizontal transfer of ATPase genes--the tree of life becomes a net of life. Biosystems. 31(2-3): 111-119.
Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol. 47(1): 3-8.
Hrdy, I., R. P. Hirt, P. Dolezal, L. Bardonova, P. G. Foster, J. Tachezy and T. M. Embley. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature. 432(7017): 618-622.
Huber, H., M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer and K. O. Stetter. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature. 417(6884): 63-67.
Huber, R., H. Huber and K. O. Stetter. 2000. Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. FEMS Microbiol Rev. 24(5): 615-623.
Huelsenbeck, J. P. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol. 19(5): 698-707.
Hull, D. L. 1980. Review: Cladism gets Sorted Out. Paleobiology. 6(1): 131-136.
Hull, D. L. 1985. Bias and Commitment in Science, Phenetics and Cladistics. Annals of Science. 42(3): 319-338.
Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa and T. Miyata. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA. 86(23): 9355-9359.
Iwasaki, W. and T. Takagi. 2007. Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics. 23(13): i230-239.
Jain, R., M. C. Rivera and J. A. Lake. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci U S A. 96(7): 3801-3806.
Jeffroy, O., H. Brinkmann, F. Delsuc and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22(4): 225-231.
Jensen, R. J. 2009. Phenetics: revolution, reform or natural consequence? Taxon. 58: 50-60.
Jones, D. T., W. R. Taylor and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8(3): 275-282.
Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules.In: H. N. Munro. Mammalian protein metabolism. New York: Academic Press**:** 21-132.
Kanhere, A. and M. Vingron. 2009. Horizontal Gene Transfers in prokaryotes show differential preferences for metabolic and translational genes. BMC Evol Biol. 9: 9.

Karner, M. B., E. F. DeLong and D. M. Karl. 2001. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature. 409(6819): 507-510.

Keeling, P. J. and J. D. Palmer. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 9(8): 605-618.

Kennedy, S. P., W. V. Ng, S. L. Salzberg, L. Hood and S. DasSarma. 2001. Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res. 11(10): 1641-1650.

Kim, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45(3): 363-374.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. Genetics. 47: 713-719.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16(2): 111-120.

Kletzin, A. 2007. General Characteristics and Important Model Organisms.In: R. Cavicchioli. Archaea, Molecular and Cellular Biology. Washington DC: ASM Press**:** 523.

Koonin, E. V., K. S. Makarova and L. Aravind. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol. 55: 709-742.

Koonin, E. V. and Y. I. Wolf. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 36(21): 6688-6719.

Korbel, J. O., B. Snel, M. A. Huynen and P. Bork. 2002. SHOT: a web server for the construction of genome phylogenies. Trends Genet. 18(3): 158-162.

Koshi, J. M. and R. A. Goldstein. 1998. Models of natural mutations including site heterogeneity. Proteins. 32(3): 289-295.

Koshi, J. M., D. P. Mindell and R. A. Goldstein. 1999. Using physical-chemistry-based substitution models in phylogenetic analyses of HIV-1 subtypes. Mol Biol Evol. 16(2): 173-179.

Koski, L. B., R. A. Morton and G. B. Golding. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol. 18(3): 404-412.

Kreil, D. P. and C. A. Ouzounis. 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. 29(7): 1608-1615.

Kupczok, A., H. A. Schmidt and A. von Haeseler. 2010. Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. Algorithms Mol Biol. 5(1): 37.

Kurland, C. G., B. Canback and O. G. Berg. 2003. Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A. 100(17): 9658-9662.

Lanave, C., G. Preparata, C. Saccone and G. Serio. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol. 20(1): 86-93.

Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin and N. R. Pace. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A. 82(20): 6955-6959.

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21(6): 1095-1109.

Lartillot, N., H. Brinkmann and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7 Suppl 1: S4.

Lartillot, N. and H. Philippe. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci. 363: 1463–1472.

Lartillot, N., T. Lepage and S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25(17): 2286-2288.

Lawrence, J. G. and H. Ochman. 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. 44(4): 383-397.

Lawrence, J. G. and H. Ochman. 2002. Reconciling the many faces of lateral gene transfer. Trends Microbiol. 10(1): 1-4.

Le, S. Q. and O. Gascuel. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25(7): 1307-1320.

Le, S. Q., N. Lartillot and O. Gascuel. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B Biol Sci. 363(1512): 3965-3976.

Lemmon, A. R., J. M. Brown, K. Stanger-Hall and E. M. Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst Biol. 58(1): 130-145.

Lercher, M. J. and C. Pal. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol. 25(3): 559-567.

Li, C., G. Orti, G. Zhang and G. Lu. 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol Biol. 7: 44.

Li, L., C. J. Stoeckert, Jr. and D. S. Roos. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9): 2178-2189.

Lin, J. and M. Gerstein. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. 10(6): 808-818.

Lio, P. and N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Res. 8(12): 1233-1244.

Liolios, K., K. Mavromatis, N. Tavernarakis and N. C. Kyrpides. 2008. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Research. 36(suppl 1): D475-D479.

Lockhart, P., M. Steel, M. Hendy and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11(4): 605-612.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland and A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. 34(2): 153-162.

Lopez-Garcia, P. and D. Moreira. 1999. Metabolic symbiosis at the origin of eukaryotes. Trends Biochem Sci. 24(3): 88-93.

Lopez, P., D. Casane and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. 19(1): 1-7.
Lorenz, F. 1987. Teaching about influence in Simple Regression. Teaching Sociology(15): 173-177.
Maddison, W. P. 1997. Gene trees in species. Syst. Biol. 46(3): 523-536.
Maddison, W. P. and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst Biol. 55(1): 21-30.
Madsen, O., M. Scally, C. J. Douady et al. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature. 409(6820): 610-614.
Mae-Wan Ho, A. R., Joe Cummins. 1999. Cauliflower Mosaic Viral Promoter - A Recipe for Disaster? Microbial Ecology in Health and Disease. 11(4): 194-197.
Makarenkov, V. 2001. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics. 17(7): 664-668.
Marri, P. R., W. Hao and G. B. Golding. 2006. Adaptive evolution: the role of laterally transferred genes. BMC Evol Biol. in press.
Martin, W. and M. Müller. 1998. The hydrogen hypothesis for the first eukaryote. Nature. 392(6671): 37-41.
Matte-Tailliez, O., C. Brochier, P. Forterre and H. Philippe. 2002. Archaeal phylogeny based on ribosomal proteins. Mol Biol Evol. 19(5): 631-639.
Miklos, I., A. Novak, R. Satija, R. Lyngso and J. Hein. 2009. Stochastic models of sequence evolution including insertion-deletion events. Stat Methods Med Res. 18(5): 453-485.
Mirkin, B., I. Muchnik and T. F. Smith. 1995. A biologically consistent model for comparing molecular phylogenies. J Comput Biol. 2(4): 493-507.
Mirkin, B. G., T. I. Fenner, M. Y. Galperin and E. V. Koonin. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol. 3(1): 2.
Miyamoto, M. M. and W. M. Fitch. 1996. Constraints on protein evolution and the age of the eubacteria/eukaryote split. Syst. Biol. 45: 566-.
Mongodin, E. F., K. E. Nelson, S. Daugherty et al. 2005. The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. Proc Natl Acad Sci U S A. 102(50): 18147-18152.
Moreno-Hagelsieb, G. and K. Latimer. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 24(3): 319-324.
Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. Nature. 409(6820): 614-618.
Musto, H., H. Naya, A. Zavala, H. Romero, F. Alvarez-ValIn and G. Bernardi. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Letters. 573(1-3): 73-77.
Musto, H., H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin and G. Bernardi. 2006. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun. 347(1): 1-3.
Nakhleh, L., D. Ruths and L.-S. Wang. 2005. RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer.In: L. Wang. Computing and Combinatorics: Springer Berlin / Heidelberg. **3595:** 84-93.

Nei, M. and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3(5): 418-426.

Nelson, K. E., R. A. Clayton, S. R. Gill et al. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature. 399(6734): 323-329.

Nishihara, H., N. Okada and M. Hasegawa. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol. 8(9): R199.

Novichkov, P. S., M. V. Omelchenko, M. S. Gelfand, A. A. Mironov, Y. I. Wolf and E. V. Koonin. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol. 186(19): 6575-6585.

Novozhilov, A. S., G. P. Karev and E. V. Koonin. 2005. Mathematical modeling of evolution of horizontally transferred genes. Mol Biol Evol. 22(8): 1721-1732.

Ochman, H., J. G. Lawrence and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405(6784): 299-304.

Ochman, H. 2001. Lateral and oblique gene transfer. Curr Opin Genet Dev. 11(6): 616-619.

Ochman, H., Lawrence, JG. 1996. Phylogenetics and the amelioration of bacterial genomes.In: F. C. N. e. al. Escherichia coli and Salmonella typhimurium : Molecular and Cellular Biology. Washington D.C.: ASM Publications.

Olendzenski, L. and J. P. Gogarten. 2009. Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer. Ann N Y Acad Sci. 1178: 137-145.

Olsen, G. J. and C. R. Woese. 1993. Ribosomal RNA: a key to phylogeny. Faseb J. 7(1): 113-123.

Oren, A. 2002. Molecular ecology of extremely halophilic Archaea and Bacteria. FEMS Microbiology Ecology. 39(1): 1-7.

Page, R. D. M. 1993. On Islands of Trees and the Efficacy of Different Methods of Branch Swapping in Finding Most-Parsimonious Trees. Systematic Biology. 42(2): 200-210.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol. 53(4): 571-581.

Pagel, M. and A. Meade. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. Philos Trans R Soc Lond B Biol Sci. 363(1512): 3955-3964.

Philippe, H. 1993. MUST, a computer package of Management Utilities for Sequences and Trees. Nucleic Acids Res. 21(22): 5264-5272.

Philippe, H. and J. Laurent. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev. 8(6): 616-623.

Philippe, H. 2000. Long branch attraction and protist phylogeny. Protist. 51(4): 307-316.

Philippe, H. and C. J. Douady. 2003. Horizontal gene transfer and phylogenetics. Curr Opin Microbiol. 6(5): 498-505.

Philippe, H., N. Lartillot and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol. 22(5): 1246-1253.

Philippe, H., F. Delsuc, H. Brinkmann and N. Lartillot. 2005. Phylogenomics. Annu Rev Ecol Evol Syst. 36: 541-562.

Philippe, H., H. Brinkmann, P. Martinez, M. Riutort and J. Baguna. 2007. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLoS ONE. 2: e717.

Philippe, H., R. Derelle, P. Lopez et al. 2009. Phylogenomics revives traditional views on deep animal relationships. Curr Biol. 19(8): 706-712.

Philippe, H., H. Brinkmann, R. R. Copley, L. L. Moroz, H. Nakano, A. J. Poustka, A. Wallberg, K. J. Peterson and M. J. Telford. 2011a. Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature. 470(7333): 255-258.

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. Littlewood, M. Manuel, G. Worheide and D. Baurain. 2011b. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9(3): e1000602.

Phillips, M. J., F. Delsuc and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21: 1455-1458.

Pisani, D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. Systematic Biology. 53(6): 978-989.

Pisani, D., J. A. Cotton and J. O. McInerney. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 24(8): 1752-1760.

Poptsova, M. 2009. Testing Phylogenetic Methods to Identify Horizontal Gene Transfer.In: M. B. Gogarten, J. P. Gogarten and L. C. Olendzenski. Horizontal Gene Transfer: Humana Press. **532:** 227-240.

Pride, D. T., R. J. Meinersmann, T. M. Wassenaar and M. J. Blaser. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Res. 13(2): 145-158.

Puigbo, P., Y. I. Wolf and E. V. Koonin. 2010. The tree and net components of prokaryote evolution. Genome Biol Evol. 2: 745-756.

Qi, J., B. Wang and B. I. Hao. 2004. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. J. Mol. Evol. 58(1): 1-11.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni et al. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature. 402(6760): 404-407.

Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees. Mol Phylogenet Evol. 1(1): 53-58.

Ragan, M. A. 2001. Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev. 11(6): 620-626.

Ragan, M. A. 2002. Reconciling the many faces of lateral gene transfer. Trends in Microbiology. 10(1): 4-4.

Ranwez, V., A. Criscuolo and E. J. Douzery. 2010. SuperTriplets: a triplet-based supertree approach to phylogenomics. Bioinformatics. 26(12): i115-123.

Ren, F., H. Tanaka and Z. Yang. 2009. A likelihood look at the supermatrix-supertree controversy. Gene. 441(1-2): 119-125.

Richardson, A. O. and J. D. Palmer. 2007. Horizontal gene transfer in plants. J Exp Bot. 58(1): 1-9.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences. 53(1-2): 131-147.

Robinson, D. M., D. T. Jones, H. Kishino, N. Goldman and J. L. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol. 20(10): 1692-1704.

Rodrigue, N., N. Lartillot, D. Bryant and H. Philippe. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 347(2): 207-217.

Rodrigue, N., H. Philippe and N. Lartillot. 2006. Assessing site-interdependent phylogenetic models of sequence evolution. Mol Biol Evol. 23(9): 1762-1775.

Rodrigues, E., M. Sagot and Y. Wakabayashi. 2007. The maximum agreement forest problem: Approximation algorithms and computational experiments. Theor Comput Sci(374): 91-110.

Rodriguez-Ezpeleta, N., H. Brinkmann, S. C. Burey, B. Roure, G. Burger, W. Loffelhardt, H. J. Bohnert, H. Philippe and B. F. Lang. 2005. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. Current Biology. 15(14): 1325-1330.

Rodriguez-Ezpeleta, N., H. Philippe, H. Brinkmann, B. Becker and M. Melkonian. 2007a. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of mesostigma in the streptophyta. Mol Biol Evol. 24(3): 723-731.

Rodriguez-Ezpeleta, N., H. Brinkmann, B. Roure, N. Lartillot, B. F. Lang and H. Philippe. 2007b. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. 56(3): 389-399.

Rokas, A., B. L. Williams, N. King and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425(6960): 798-804.

Rota-Stabelli, O., L. Campbell, H. Brinkmann, G. D. Edgecombe, S. J. Longhorn, K. J. Peterson, D. Pisani, H. Philippe and M. J. Telford. 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc Biol Sci. 278: 298-306.

Roure, B., N. Rodriguez-Ezpeleta and H. Philippe. 2007. SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. BMC Evol Biol. 7 Suppl 1: S2.

Roure, B. and H. Philippe. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evol Biol. 11(1): 17.

Schleper, C., R. V. Swanson, E. J. Mathur and E. F. DeLong. 1997. Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon Cenarchaeum symbiosum. J Bacteriol. 179(24): 7803-7811.

Slesarev, A. I., K. V. Mezhevaya, K. S. Makarova et al. 2002. The complete genome of hyperthermophile *Methanopyrus kandleri AV19* and monophyly of archaeal methanogens. Proc Natl Acad Sci U S A. 99(7): 4644-4649.

Snel, B., P. Bork and M. A. Huynen. 1999. Genome phylogeny based on gene content. Nat Genet. 21(1): 108-110.

Snel, B., P. Bork and M. A. Huynen. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 12(1): 17-25.

Soltis, D. E., V. A. Albert, V. Savolainen et al. 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci. 9(10): 477-483.

Soltis, P. S., D. E. Soltis and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature. 402(6760): 402-404.

Sonnhammer, E. L. and E. V. Koonin. 2002. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. 18(12): 619-620.

Spang, A., R. Hatzenpichler, C. Brochier-Armanet et al. 2010. Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. Trends Microbiol. 18(8): 331-340.

Sperling, E. A., K. J. Peterson and D. Pisani. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Mol Biol Evol. 26(10): 2261-2274.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22(21): 2688-2690.

Stefanovic, S., D. W. Rice and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? BMC Evol Biol. 4(1): 35.

Stetter, K. O. 1996. Hyperthermophilic procaryotes. FEMS Microbiology Reviews. 18(2-3): 149-158.

Suchard, M. A. 2005. Stochastic models for horizontal gene transfer: taking a random walk through tree space. Genetics. 170(1): 419-431.

Susko, E. and A. J. Roger. 2007. On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol. 24(9): 2139-2150.

Talavera, G. and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56(4): 564-577.

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10(3): 512-526.

Team, R. D. C. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Tekaia, F., A. Lazcano and B. Dujon. 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. 9(6): 550-557.

Thomas, C. M. and K. M. Nielsen. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol. 3(9): 711-721.

Thorne, J. L., N. Goldman and D. T. Jones. 1996. Combining protein evolution and secondary structure. Mol Biol Evol. 13(5): 666-673.

Tillier, E. and R. Collins. 1995. Neighbor Joining and Maximum Likelihood with RNA Sequences: Addressing the Interdependence of Sites. Molecular Biology and Evolution. 12(1): 7.

Trivedi, S., H. S. Gehlot and S. R. Rao. 2006. Protein thermostability in Archaea and Eubacteria. Genet Mol Res. 5(4): 816-827.

Tuffley, C. and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci. 147(1): 63-91.

Valentine, D. L. 2007. Adaptations to energy stress dictate the ecology and evolution of the Archaea. Nat Rev Microbiol. 5(4): 316-323.

Vetriani, C., A. L. Reysenbach and J. Dore. 1998. Recovery and phylogenetic analysis of archaeal rRNA sequences from continental shelf sediments. FEMS Microbiol Lett. 161(1): 83-88.

Vinh le, S. and A. Von Haeseler. 2004. IQPNNI: moving fast through tree space and stopping in time. Mol Biol Evol. 21(8): 1565-1571.

Wang, B. 2001. Limitations of compositional approach to identifying horizontally transferred genes. J Mol Evol. 53(3): 244-250.
Wang, H. C., M. Spencer, E. Susko and A. J. Roger. 2007. Testing for covarion-like evolution in protein sequences. Mol Biol Evol. 24(1): 294-305.
Wang, H. C., K. Li, E. Susko and A. J. Roger. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol. 8: 331.
Waterman, M. S. and T. F. Smith. 1978. On the similarity of dendrograms. Journal of Theoretical Biology. 73(4): 789-800.
Waters, E., M. J. Hohn, I. Ahel et al. 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A. 100(22): 12984-12988.
Welch, R. A., V. Burland, G. Plunkett, 3rd et al. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A. 99(26): 17020-17024.
Whelan, S. and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18(5): 691-699.
Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst Biol. 54(5): 731-742.
Williams, D. and M. Ebach. 2009. What, Exactly, is Cladistics? Re-writing the History of Systematics and Biogeography. Acta Biotheoretica. 57(1): 249-268.
Woese, C. R. and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA. 74(11): 5088-5090.
Woese, C. R. 1987. Bacterial evolution. Microbiol Rev. 51(2): 221-271.
Woese, C. R., O. Kandler and M. L. Wheelis. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA. 87(12): 4576-4579.
Woese, C. R., L. Achenbach, P. Rouviere and L. Mandelco. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst Appl Microbiol. 14(4): 364-371.
Wolf, Y. I., I. B. Rogozin, N. V. Grishin, R. L. Tatusov and E. V. Koonin. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol. 1(1): 8.
Wolf, Y. I., I. B. Rogozin, N. V. Grishin and E. V. Koonin. 2002. Genome trees and the tree of life. Trends Genet. 18(9): 472-479.
Wuchter, C., B. Abbas, M. J. Coolen et al. 2006. Archaeal nitrification in the ocean. Proc Natl Acad Sci U S A. 103(33): 12317-12322.
Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol. 10(6): 1396-1401.
Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39(3): 306-314.
Yang, Z. and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol Biol Evol. 12(3): 451-458.
Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42: 587-596.

Zuckerkandl, E. and L. Pauling. 1965. Molecules as documents of evolutionary history. J Theor Biol. 8(2): 357-366.
Zwickl, D. J. and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51(4): 588-598.

Annexe 1 - Protocole élaboré dans le cadre de l'étude des Archaea

Dans les paragraphes qui suivront vous seront énumérées les différentes étapes constituant le protocole de découverte de gènes orthologues putatifs, d'inférence d'arbre phylogénétique ainsi que de la détection de transferts horizontaux de gènes parmi les gènes orthologues putatifs pour un jeu de données contenant les protéomes complets des organismes visés. L'étape de détection est immédiatement suivie d'une filtration de ces alignements.

Recherche des gènes orthologues putatifs

La première étape de ce protocole consiste à télécharger et trouver les différents protéomes complets des organismes que nous voulons étudier. Ceci peut se faire en allant directement sur le site FTP du NCBI, où tous les génomes et protéomes complets sont disponibles. Les protéomes sont disponibles dans les fichiers « *.faa* » sur le site du **NCBI**. Il faut télécharger et enregistrer les différents protéomes dans des fichiers différents pour que l'outil de détection des orthologues sache combien d'espèces sont présentes dans les regroupements qu'il fera. Par la suite, la recherche de gènes orthologues peut débuter. Pour cela, nous utilisons le programme **OrthoMCL** (Li, Stoeckert et Roos, 2003), qui utilise l'outil de recherche de similarité de séquence **BLAST** (Altschul et al., 1990) afin de faire la recherche des orthologues. Il faut que les deux soient bien installés sur le système, en plus d'un autre programme requis par **OrthoMCL**, **mcl**, qui est un algorithme de regroupement markovien (markov clustering). Des paramètres de **BLAST** ont été utilisés afin d'optimiser cette recherche d'orthologues (Moreno-Hagelsieb et Latimer, 2008).

Une fois le programme installé, il faut regrouper tous les protéomes au même endroit, soit dans un sous-dossier d'où se situe **OrthoMCL** (la version la plus récente, 2.0, utilise des bases de données MySQL). Cependant, pour nos besoins, nous ferons le travail avec l'ancienne version, soit la **version 1.4**. Il est recommandé de lancer ce processus sur une machine possédant plusieurs processeurs afin de paralléliser le plus possible l'étape ou le programme fait le **BLAST** tous contre tous. **OrthoMCL** peut se lancer sous plusieurs

modes. Lorsque le jeu semble raisonnable (ex. de petits génomes (bactéries) en quantité inférieure à 80) nous pouvons lancer avec le mode 1, qui effectue toutes les étapes pour le regroupement. Par contre, lorsque le jeu de données est très grand, il est possible qu'**OrthoMCL**, qui est codé en Perl, ait des problèmes de mémoire lorsqu'il passe d'un programme à l'autre (**BLAST** vers **mcl**). Il faudra donc vérifier le tout et si la mémoire flanche, lancer avec les fichiers générés par **BLAST**, sur le mode 3.

OrthoMCL génèrera comme sortie finale un simple fichier texte avec les différents regroupements, le nom des gènes ainsi que le nom des espèces impliquées. Il faudra alors être certain que les noms des espèces correspondent aux noms des fichiers des différents protéomes pour la prochaine étape. Cette étape implique un script que j'ai codé. Il consiste à construire les regroupements en ayant seulement le fichier de sortie d'**OrthoMCL** et les différents fichiers de protéomes. Ce script se nomme **oMCL-fastaExtractor_gene.pl** . Il faut aller changer dans le script, avant de le lancer, le dossier de sortie ainsi que le dossier où les différents protéomes sont situés. L'entrée est le fichier de sortie d'**OrthoMCL**. Ceci nous créera les différents regroupements sous format *fasta*. De plus, dans ce script, il est possible de mettre un seuil (identifié *$geneThreshold*) qui nous permet de dire si nous voulons les regroupements avec par exemple plus de deux fois le nombre de séquences que d'espèces. Ceci a été implanté puisque certains regroupements peuvent contenir par exemple 200 séquences et seulement 10 espèces, ce qui n'est pas très utile et serait très difficile à analyser. Cette option est désactivée dans la version présente.

Changement de format des fichiers de gènes orthologues

Les prochaines étapes sont interchangeables. Nous avons désormais les différents fichiers de gènes orthologues putatifs sous format *fasta*. Cependant, il faut maintenant s'assurer que les entêtes des fichiers de gènes sont uniformes afin de bien travailler avec ceux-ci par la suite. La plupart du temps, si les fichiers ont été téléchargés de la base de données du NCBI, le fichier sera formaté de la façon suivante :

gi#######|xxxxxxxx Description [Espèce]

Un script a également été écrit afin de pouvoir transformer ces fichiers sous un bon format qui pourra être utilisé par la suite dans les différents autres programmes. Il s'agit du script **fastaHeaderChange.pl**. Il suffit de le lancer en boucle sur tous les fichiers « *.cl »* résultant de l'étape précédente. Les fichiers de sortie seront des fichiers « *.fasta2* ». Vous pourrez par la suite les renommer en « *.fasta* ». Le nom d'extension du fichier de sortie peut également être changé dans le script facilement.

Sélection des gènes

L'une des étapes les plus importantes au départ est de sélectionner le seuil pour les regroupements à sélectionner. **OrthoMCL** classe les différents regroupements dans l'ordre du nombre de séquences, et non pas du nombre d'espèces. Par exemple, une tendance fortement observée est que le regroupement #0 (le premier dans la liste) a un très grand nombre de séquences, mais très peu d'espèces. C'est pour cela qu'il faudra sélectionner les gènes qui ont un nombre d'espèces suffisant comparé au nombre de séquences. Par exemple, pour un jeu de données testé, j'ai sélectionné tous les regroupements ayant au moins 50 espèces sur 79, ce qui fait au plus 37% d'espèces manquantes. Si nous voulions être moins sévères, nous aurions pu accepter plus d'espèces manquantes. Afin de trouver cela, la commande **seq** de Linux a été utilisée afin de trouver les regroupements avec plus de 50 espèces.

```
%> for f in `seq 50 1 79` ; do cp *_$f.fasta ../Archaea_50_79sp ; done
```

Il faut par la suite regrouper tous ces fichiers dans un autre dossier et nous serons prêts pour la prochaine étape.

Alignement et filtration des positions conservées

La prochaine étape est cruciale pour le type d'analyse qui se fera par la suite. Il s'agit d'aligner les séquences, protéiques dans notre cas. Pour cela, vous pouvez utiliser un programme d'alignement prenant en entrée le format *fasta*. **Muscle** (Edgar, 2004) est recommandé pour des raisons de rapidité et de très bonne performance dans l'alignement. Par la suite, avec les données format aligné, une application de bornes se fera à l'aide de **GBlocks** (Talavera et Castresana, 2007). Ce programme permet de conserver seulement les « blocs » de positions conservées au cours de l'évolution selon

différents seuils. Les paramètres par défaut ont été changés afin d'avoir un plus grand nombre de positions par gènes tout en gardant une certaine stringence.

Nombre maximum de positions contigües non-conservées :5
Longueur minimum pour un bloc : 2
Sauts ("Gap") acceptés : h (pour moitié des séquences)

Une fois tout cela complété, les séquences, encore en format *fasta*, peuvent être transformées en format *ali*, le format utilisé pour la majorité des applications dans le laboratoire (Philippe, 1993).

Concaténations

La prochaine étape permettra de faire le jeu de référence pour le reste des manipulations. Il a été démontré selon différentes mesures qu'une matrice de distances de référence comportant toutes les séquences à l'étude restait très robuste aux effets liés aux transferts horizontaux.

Il faudra faire la concaténation de toutes les séquences à l'étude. Le programme **SCaFoS** (Roure, Rodriguez-Ezpeleta et Philippe, 2007) peut être utilisé afin de faire cette étape. De plus, différentes options s'offrent à vous afin de pouvoir sélectionner une copie de paralogues présents dans votre regroupement d'orthologues putatifs. Le choix de la distance minimale d'évolution semble le meilleur pour ce cas-ci. Vous aurez après cette étape votre concaténation de référence. Il faut s'assurer à cette étape que les noms des séquences soient **triés en ordre alphabétique** dans notre fichier d'alignement obtenu. Ce sera d'une importance primordiale pour le reste du protocole. Vous aurez le fichier d'alignement sous format *fasta* après cette étape. Afin de pouvoir effectuer la prochaine étape, vous devrez convertir ce fichier en format « *pseudo-phylip* » avec le script **fasta2relaxedphylip.pl**. Ce script sera aussi utilisé plus tard par un autre script et nous permettra d'obtenir l'alignement de séquences dans le bon format en plus de ne pas avoir la contrainte de 10 caractères pour l'identifiant des séquences, contrainte existante dans le format phylip conventionnel. Il sera alors important de l'avoir dans votre dossier *bin* sur votre racine afin d'y avoir accès, tout cela est également valable pour tous les autres scripts qui seront décrits plus loin.

Les informations qui suivent de ce protocole sont également élaborées et mieux explicitées dans l'Annexe 2.

Calcul de la matrice de distances de référence

L'étape suivante nécessite, tout dépendant de la taille de votre jeu de données, l'utilisation de machine ayant à disponibilité beaucoup de mémoire vive. Par exemple, un jeu d'environ 90 000 positions avec 79 espèces demandera plus de 12 Go de mémoire. La présente étape est le calcul d'une matrice de distance corrigée sous la distribution de vitesse d'évolution Gamma avec 4 catégories avec le modèle d'évolution WAG (Whelan et Goldman, 2001). Vous pourrez faire le tout avec le programme **IQPNNI** (Vinh le et Von Haeseler, 2004).

%> iqpnni64 –prefix job –m WAG –w gamma –c 4 –n 0 –param param.txt job.phylip

Mettre dans *param.txt* seulement *y*. De cette façon, il répondra automatiquement « y » à la question qu'il posera. Récupérez par la suite le fichier *job.iqpnni.dist* et renommez-le en *job.mat*.

But de l'approche

L'approche consistera à évaluer individuellement chaque séquence dans un regroupement de gènes orthologues. Cette évaluation se fera à partir de matrices de distances corrigées calculées avec **IQPNNI**. Il s'agira d'évaluer la variation du coefficient de corrélation entre la matrice de distances de référence et la matrice de distances des gènes individuels lorsque nous enlevons et remettons une séquence à la fois.

Établissement de la courbe des seuils selon la longueur des séquences

****Tout ce qui suit est automatisé et peut être lancé avec la commande :*

%> threshold_correlation_calc.pl --referencefile=<ref.fasta> --matfile=<matfile.mat> [--rep=NbReplicats]

Vous n'avez qu'à vous créer un nouveau dossier ne contenant que le fichier de matrice de distances et le fichier d'alignement de référence (avec les identifiants en ordre alphabétique).

Cette étape consiste à déterminer le seuil duquel une séquence sera évaluée. En fait, une séquence sera enlevée définitivement si, lorsque nous la retirons de la matrice de distances, le coefficient de corrélation grimpe d'une valeur supérieure au seuil prédéterminé par la courbe lié à la taille. L'étape cruciale pour cette filtration sera alors de déterminer cette courbe de seuils.

Afin de calculer celle-ci, nous devons tirer aléatoirement des positions dans notre concaténation de référence, de façon à bâtir des jeux de différentes tailles, ayant une composition neutre et de même composition que le jeu de référence. Il s'agit d'un processus de tirage par « jacknife ». Un programme existe afin d'effectuer cette opération. Il s'agit de **jacksite***,* et il ne faut lui fournir en entrée que le jeu dans lequel nous voulons tirer les positions, le nombre de positions et le nombre de réplicats. Afin d'avoir une bonne moyenne pour chaque point mesuré, 100 réplicats sont de mise. Cette étape demande une structure impeccable dans les dossiers, car le programme donne en sortie des noms de fichiers identiques à chaque fois que celui-ci est lancé. Il faut alors renommer les fichiers de sortie toutes les fois pour que ceux-ci aient une signification : ex. *seqdata1_50pos.phylip*. Un autre problème existe dans ce cas également. Il faut convertir les fichiers en format *phylip* vers le format *ali* afin de pouvoir être en mesure d'effectuer les comparaisons de matrices pour cette étape. Le problème réside dans le fait que les données manquantes (gap ou autre) sont transformées en « espace » (« ») littéralement par le script de conversion **puz2ali**. Il faut alors, après la conversion, changer les espaces par des « ? ».

Pour mesurer les points de seuil, il faudra, pour chaque jeu construit par jacknife, faire tourner le programme **correlation_coeff_calcV3.pl** et simplement regarder les valeurs de variation obtenues dans les fichiers « .results ». Vous devrez enlever les quatre premières lignes avec la commande :

```
%> sed '1,4d' Reference_mapped_seqdata1_50pos.results >
 Reference_mapped_seqdata1_50pos.results.out
```

(si la matrice de distance de référence se nommait Reference.mat).

****Rappel tout ceci est automatisé****

Afin de bien exécuter le script **correlation_coeff_calcV3.pl**, créez-vous un dossier contenant tous vos fichiers *ali* que vous venez de générer, ainsi que le fichier de matrice de référence que vous avez calculé plus haut. Le script se lance de cette façon :

%> correlation_coeff_calcV3.pl --inputfile=<inputfile.ali> --referencefile=<ref.mat>

Par la suite, mettez les uns à la suite des autres, pour chaque catégorie de longueur différente, tous les résultats obtenus. Vous pourrez obtenir les statistiques désirées avec le programme de statistiques **R**. Vous aurez à trouver les intervalles de confiance à 99% des données simulées. De cette façon, puisque les jeux sont assez uniformes, nous ne rejetterons que les données qui ressortent de la majorité de nos données. Le but par la suite sera d'aller chercher le point maximal de toutes ces données, soit la borne positive. De cette façon, nous saurons de quelle façon une séquence dans un jeu totalement neutre fait varier au maximum le coefficient de corrélation.

Votre fichier d'entrée sera alors un fichier avec 3 colonnes, soit le nom de l'espèce retirée, la nouvelle valeur du coefficient de corrélation entre les matrices de distances et la variation en pour cent (%). Les valeurs considérées seront celles du pourcentage.

Les catégories suggérées afin d'avoir une courbe avec suffisamment de précision sont des alignements de 50, 100, 150, 200, 250, 300, 500, 750 et 1000 positions. Cela fait beaucoup de jeux à créer.

****À partir d'ici, vous pourrez travailler avec les résultats générés par le script d'automatisation.

**** Les résultats de cette étape seront dans le fichier *results_confidence/ref_all_99.out*. Vous n'aurez par la suite qu'à copier les valeurs contenues dans ce fichier dans **Excel** ou tout autre programme et effectuer une courbe exponentielle sur ces valeurs comme ci-dessous.

Bien sûr, dans nos données réelles, il est très rare qu'aucune tendance ne soit observée et la courbe de seuil sera beaucoup trop sévère. En outre, si l'échelle de la courbe reste inchangée, nous risquons de conserver seulement les séquences qui ont un effet nul ou positif sur le coefficient de corrélation avec la matrice de référence. Il faut donc changer

l'échelle de cette courbe et après plusieurs tests sur quelques jeux différents, une augmentation d'un facteur 100 semble être un bon compromis. Voici ce à quoi une telle courbe devrait avoir l'air avant et après le changement d'échelle.

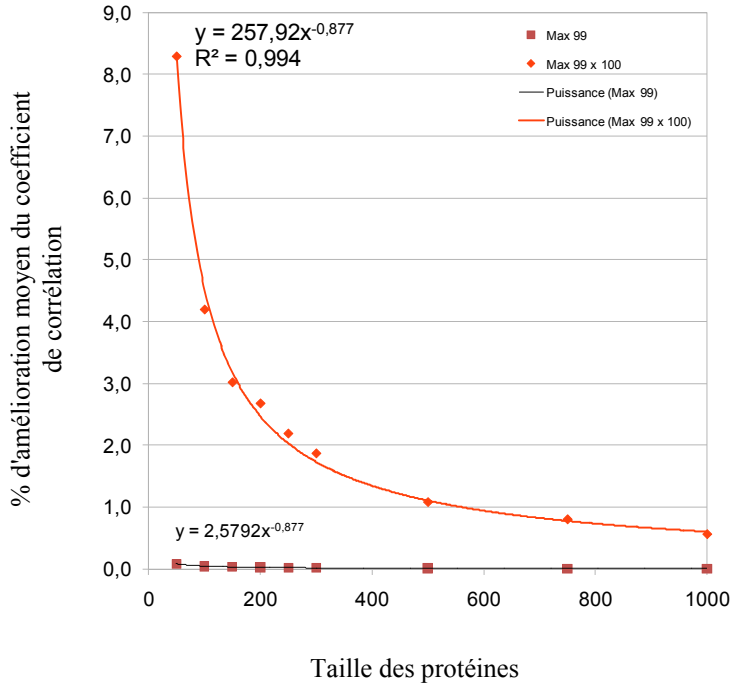


Figure A1.1 Courbe de seuils pour le programme de filtration d'alignement de séquences. Le pourcentage d'amélioration du coefficient de corrélation représente l'augmentation de celui-ci lors du retrait d'une séquence mauvaise.

Vous pouvez voir, selon les formules, que le changement d'échelle ne change pas l'allure de la courbe. Cette courbe, en rouge, signifie alors que nous retirons une séquence d'une taille de 150 sites, par exemple, seulement si lorsque nous la retirons, le coefficient de corrélation s'améliore de plus de 3%. Lorsqu'il s'agira de choisir entre plusieurs paralogues, il s'agira d'une petite subtilité dans l'utilisation de cette courbe. Vous devrez donc ajouter la valeur de cette courbe à la fin du script **correlation_coeff_calcV3.pl**.

Analyse des données réelles et filtration des alignements

Nous sommes maintenant rendus à l'étape finale de ce protocole. Il s'agit d'exécuter le script d'analyse des coefficients de corrélation sur les données réelles. Pour cela, veuillez vous créer un dossier comprenant tous vos fichiers de regroupements d'orthologues

putatifs ainsi que la matrice de référence. Vous pouvez par la suite lancer en boucle **BASH** le script sur tous les fichiers.

```
%> for i in *.ali ; do correlation_coeff_calcV2_3.pl --inputfile=$i --referencefile=ref.mat ; done
```

Lorsque tout est terminé, le script crée un dossier *fichier_Analysis* (où le fichier en entrée était *fichier.ali*). Dans ce dossier, plusieurs types de fichiers pourront servir à faire les concaténations désirées. Voici la description de ces fichiers :

- fichier.ali #Fichier d'alignement initial remis en ordre alphabétique
- fichier.mat #Matrice de distances corrigée calculée sur l'alignement initial
- fichier_XXsp_XXpos.ali #Fichier d'alignement filtré final
- fichier_XXsp_XXpos_int.ali #Fichier d'alignement avec le meilleur paralogue (non-filtré)
- fichier_XXsp_XXpos_final.mat #Matrice de distance finale filtrée
- fichier_XXsp_XXpos_int.mat #Matrice de distance intermédiaire (avec le meilleur paralogue)
- reference.mat #Matrice de distance de référence initiale
- reference_mapped.mat #Matrice de distance de référence sans les espèces non présentes
- reference_mapped_int.mat #Matrice de distance de référence intermédiaire
- reference_mapped_final.mat #Matrice de distance de référence finale filtrée
- reference_mapped_int_VS_fichier_XXsp_XXpos_int.coefInter #Coefficient de corrélation intermédiaire
- reference_mapped_int_VS_fichier_XXsp_XXpos_int.lmInter #Coefficient de corrélation final, après filtr.
- Cas #1 : Aucun paralogue
    - reference_mapped_fichier.results #Fichier contenant la variation du coefficient de corrélation pour chaque séquence
- Cas #2 : Présence de paralogues
    - reference_mapped_fichier.results_core #Fichier contenant les variations des coefficients de corrélation pour le cœur de séquences, celles n'ayant pas de paralogues
    - reference_mapped_fichier.dupresults #Fichier contenant les variations de coefficient de corrélation pour chaque copie des paralogues ainsi que les séquences ayant été enlevées du cœur de séquences

mapped : Élimination des séquences non présentes dans le fichier de gène individuel de la référence
int : État intermédiaire, sans les mauvais paralogues

Les fichiers contenant les valeurs de coefficients de corrélation avant et après la filtration peuvent être d'une très grande utilité si nous désirons faire des concaténations selon ces critères.

Voici les programmes, scripts que vous aurez besoin dans votre *bin* :

- Fasta2Ali.pl
- oMCL-fastaExtractor_gene.pl

- fastaHeaderChange.pl
- threshold_correlation_calc.pl
- set_numbers.pl
- set_names_in_ali.pl
- ali2namepuz
- set_names_in_puzalign.pl
- iqpnni64
- correlation_coeff_calc_V2_3.pl

Annexe 2 - Protocole d'élimination des séquences possiblement problématiques

L'idée de cette approche est très simple. Trouver un moyen pour discriminer des séquences d'un gène ayant un taux d'évolution trop rapide ou trop lent par rapport à l'ensemble des autres séquences de l'espèce en question. Notre approche se veut aussi être une façon de trouver et d'éliminer des séquences qui pourraient soient être des gènes paralogues ou une séquence ayant subi un évènement de transfert horizontal. Les paragraphes qui suivront exposeront l'ensemble de cette approche, ainsi que les fondements qui ont conduit à celle-ci.

L'idée principale de notre approche était de trouver une façon d'identifier des séquences qui sortaient de la tendance générale du génome des espèces. Nous avons basé l'ensemble de notre protocole sur la comparaison de matrices de distances (Philippe, 1993). Cette idée a également été amenée par différents autres auteurs (Novichkov et al., 2004; Kanhere et Vingron, 2009), mais ceux-ci ne nous ont pas influencés lors de l'élaboration de notre protocole, puisque les articles ont été lus après que tout ait été implémenté. Le critère d'évaluation des séquences repose cependant sur un concept différent des deux autres articles, soit sur la variation du coefficient de corrélation de Pearson lors de la délétion d'une séquence.

Fichiers d'entrée

Il nous faut en entrée seulement deux fichiers. Le premier correspond à la matrice de distances de la concaténation de tous les gènes à l'étude (matrice de référence). Plusieurs mesures ont été effectuées afin de vérifier la robustesse de ces matrices face aux évènements de transferts horizontaux et aux autres types de séquences problématiques. Une première comparaison a été faite en coupant l'alignement de départ, correspondant à la concaténation de tous les gènes, en deux et en vérifiant la corrélation entre les deux matrices résultant de ces deux nouveaux alignements. La corrélation entre ces deux matrices dépassait 0.97, ce qui représente une bonne corrélation. Ceci signifie alors qu'une grande taille d'alignement est suffisante afin de supprimer le bruit lié aux différentes particularités des alignements individuels. Cette matrice servira de référence

afin de découvrir des séquences dans nos gènes qui ne seraient pas orthologues ou alors pouvant être sujettes à causer des problèmes de reconstruction. La matrice de référence peut être vue comme la tendance moyenne du protéome des espèces. Celle-ci se calcule à l'aide du programme **IQPNNI** (Vinh le et Von Haeseler, 2004) et peut nécessiter beaucoup de mémoire vive (plus de 8 Go) si la matrice à calculer provient d'alignements très grands. Le second fichier nécessaire est le fichier d'alignement du gène à vérifier en format *ali* (format *fasta* avec un entête composé de deux lignes commençant par « # »).

Le script se lance de la façon suivante :

%> **correlation_coeff_calcV3.pl** --inputfile=<inputfile.ali> --referencefile=<referencefile.mat> [--quick=no|yes] [--quickVal=1.%%] [--inputfile_NEW=<inputfile_NBsp_NBpos.ali>] [--method=pearson|kendall|spearman] [--dir=<directory>] [--threshold=value]

En plus de la fonctionnalité de base du script, qui est de découvrir, et d'enlever les séquences orthologues hors normes et paralogues d'un jeu de données, d'autres options y ont été ajoutées. Le script reconnait entre autres si la matrice de distances pour le gène à tester a déjà été calculée et ne la recalcule pas si tel est le cas. Tel qu'écrit dans la ligne de commande pour lancer le script, il y a les deux paramètres obligatoires qui représentent les deux fichiers à fournir en entrée. Entre crochets figurent par la suite les paramètres optionnels. Le paramètre « *quick* » peut être activé conjointement avec le paramètre « *quickVal* » afin d'éliminer des séquences rapides qui pourraient ne pas avoir été détectées à l'aide des options de bases. Ceci nous permet d'éliminer une séquence d'une espèce étant, par exemple, 15% plus rapide que l'ensemble de son génome (*quickVal*=1.15). Le paramètre « *inputfile_NEW* » nous permet de lancer le script pour seulement obtenir la valeur du coefficient de corrélation obtenue pour un fichier filtré. Il est également possible de sélectionner l'un des trois coefficients de corrélation disponibles dans la suite **R** (Team, 2010) avec le paramètre « *method* ». Si nous ne voulons pas envoyer les fichiers de sortie dans le dossier créé par défaut (nom du fichier de gène sans l'extension), un autre nom de dossier peut être spécifié avec l'option « *dir* » et finalement, si nous désirons mettre une valeur fixe pour le paramètre du seuil de différence de coefficient de corrélation à ne pas dépasser pour éliminer une séquence, il est possible de le faire avec le paramètre optionnel « *threshold* ».

Finalement, une ligne contenant la formule pour le calcul du seuil à utiliser selon la taille de la protéine est à ajouter à la fin du script sous la forme d'une courbe exponentielle.

Détermination du seuil correspondant à une séquence conforme à la tendance du génome

Afin de filtrer les alignements de séquences, une courbe de seuil selon la taille des protéines pour chaque jeu de données différent doit être calculée. Un script a été fait afin d'automatiser cette étape, celui-ci est expliqué dans le paragraphe qui suit.

%>**threshold_coeff_calc.pl** --referencefile=<referencefile.fasta>
--matfile=<matfile.mat> [--rep=NB_of_replicates]

Le concept élaboré consiste à tirer un lot de séquences aléatoires provenant de la concaténation de tous les gènes du jeu de données. De cette façon, des séquences de différentes tailles contenant la composition en acides aminés du jeu de données seront tirées et utilisées comme séquences considérées neutres (la moyenne des positions des séquences n'étant ni rapides et ni transférées horizontalement). Il faut alors utiliser ces séquences à travers notre approche afin d'avoir des valeurs de variation des coefficients de corrélation. Cette variation est la valeur utilisée afin d'identifier une séquence à enlever selon la taille de cette protéine. Son utilisation sera expliquée plus loin. Nous obtenons de cette façon la courbe neutre pour des séquences au contenu aléatoire ayant des tailles allant de 50 acides aminés à 1000 acides aminés. La courbe de seuils obtenue sera mise à une échelle plus pratique, soit augmentée d'un facteur 100, afin de réussir à identifier les séquences isolées dans les données réelles. Ce facteur fut déterminé à partir de plusieurs essais sur différents jeux de données réelles. Celui-ci peut être modifié selon la rigueur que nous voulons pour notre approche. Le protocole détaillé de son utilisation est disponible dans l'annexe 2.

Fonctionnement de l'approche

Deux cas peuvent survenir lorsque nous faisons la recherche de séquences isolées. Il peut s'agir d'un gène n'ayant aucune duplication, le cas le plus simple, ou alors d'un gène ayant une ou plusieurs duplications de séquences. Ces deux cas seront traités

différemment. L'idée, telle qu'expliquée plus haut, est de voir la variation du coefficient de corrélation de la comparaison de la matrice de distances de référence avec la matrice de distances du gène à vérifier. Nous commençons alors par calculer cette matrice de distances pour notre gène. Par la suite, la matrice de référence est adaptée au nombre d'espèces présentes dans l'alignement du gène. Cette étape est primordiale, car le tout est fait sur **R** (Team, 2010) qui nécessite un tel format.

Pour le premier cas, où il n'y a pas de duplication, voici les différentes étapes. Le test consiste à premièrement prendre la valeur du coefficient de corrélation entre les deux matrices. Par la suite, il suffit d'enlever une après l'autre, en les remettant par la suite, les différentes séquences. Nous voyons de cette façon l'impact qu'a une séquence pour la corrélation entre les deux matrices. Les séquences ne respectant pas le seuil calculé plus tôt, soient celles qui font diminuer trop fortement le coefficient de corrélation, seront éliminées.

Dans le deuxième cas, un cœur de séquences d'espèce n'ayant pas de gènes dupliqués sera déterminé. Pour ce cœur, nous procédons de la même façon qu'expliquée dans le paragraphe ci-dessus. Par la suite, une fois cette étape terminée, nous pouvons procéder à l'évaluation des séquences dupliquées. Celles-ci seront ajoutées l'une après l'autre et nous vérifierons si la séquence fait augmenter le coefficient de corrélation plus que le seuil qui a été fixé. Si oui, la meilleure séquence de l'espèce concernée sera conservée, sinon toutes les séquences dupliquées de l'espèce seront enlevées.

Retrait des séquences divergentes

À la suite de plusieurs observations que nous fîmes à partir de simulations, nous nous sommes aperçus que certains gènes comportant des séquences très rapides réagissaient étrangement face au critère du coefficient de corrélation. Les séquences très rapides faisaient augmenter à tort le coefficient de corrélation, rendant des gènes ayant de fausses topologies très corrélées à la matrice de référence. Une option supplémentaire a donc été ajoutée au programme. Celle-ci traite les gènes par rapport à la vitesse d'évolution des séquences en premier lieu, avant d'évaluer les coefficients de corrélation de chacune des séquences. Un seuil arbitraire est alors imposé, par exemple 115%, qui correspond à une vitesse 15% supérieure à la matrice de référence pour une espèce.

Cette vitesse est évaluée à partir des matrices de distances. Il s'agit en fait de la moyenne des distances pour une espèce envers toutes les autres présentes dans l'alignement. Ceci constitue une bonne estimation de la vitesse d'évolution de l'espèce en question. Ces vitesses sont alors comparées pour la matrice de référence avec la matrice de distances du gène à vérifier.

Vérification de la performance de l'approche sur des simulations de THG

Afin de vérifier la puissance de détection de cet outil, différentes simulations de THG (voir chapitre 2 pour plus de détails sur la méthode), qui furent utilisées pour d'autres manipulations également, ont été utilisées. Le script de base, conçu par Nicolas Galtier (Galtier, 2007), a été modifié afin de pouvoir suivre le déroulement de l'algorithme et d'ainsi obtenir un fichier contenant les différents évènements de transferts survenus pour les différents gènes simulés. Par la suite, un script a été conçu afin d'analyser tous les fichiers résultants de l'étape de filtration des séquences. Les séquences retirées devraient correspondre à des séquences ayant subi des THG ou des séquences trop rapides ou trop lentes par rapport à l'ensemble du génome de l'espèce en question. Le script doit être lancé au niveau où se trouve le dossier où figurent le dossier des fichiers d'alignement originaux (placés dans un dossier nommé 'ali_files') ainsi que le dossier dans lequel tous les fichiers provenant de l'étape de filtration sont présents. Celui-ci se lance de la façon suivante :

%> **hgt_simul_detect.pl** <Folder> <Analysis_Folder> <#Cycle>

où « *Folder* » représente le dossier global où sont placés tous les fichiers, où « *Analysis_Folder* » représente le dossier où les fichiers résultants de l'étape de filtration (celui-ci est un sous-dossier du dossier mis dans le paramètre « *Folder* »). Le paramètre « *#Cycle* » correspond au nombre de cycles qui ont été effectués par le programme de filtration. Si celui-ci n'a été effectué qu'une seule fois sur les données, le chiffre *1* sera alors donné. Cette option a été ajoutée afin de vérifier si l'application du filtre en boucle pourrait avoir un grand effet sur le taux de détection de THG simulés.

Premièrement, les informations provenant directement du programme de simulations sont assimilées. Le fichier *HGT_simul.moves* présent dans le dossier « *Folder* » est lu et

pour chaque gène simulé, un fichier est créé et mis dans le sous-dossier *hgt_files* venant tout juste d'être créé. Ces informations correspondent à tous les évènements créés par le programme de transferts.

Le script vérifie ensuite dans chaque sous-dossier d'analyse présent dans le dossier « *Analysis_Folder* » les résultats obtenus par le programme de filtration. Celui-ci copie les fichiers « *.results* » dans un nouveau sous-dossier nommé *results_files*, placé dans le dossier *Analysis_Folder* et appelle par la suite un autre script, soit le script **correlation_get_accepted.pl** qui prend en entrée les fichiers *.results*. Les fichiers sont analysés et deux fichiers sont obtenus, soient un fichier *.accepted_seq* et un fichier *.rejected_seq*, où les séquences présumément transférées figurent dans le fichier *.rejected_seq*. Un autre sous-dossier est ensuite créé, *new_ali*, où tous les fichiers filtrés sont copiés.

La comparaison des fichiers présents dans le dossier « *Folder/Analysis_Folder/results_files* » (détection) avec les fichiers présents dans le dossier « *Folder/hgt_files* » (informations des simulations) a ensuite lieu. Il sera question de la détection de la spécificité et de la sensibilité de la méthode par rapport à la détection des THG dans les alignements de séquences.

Vrai positif : Séquence affectée détectée comme THG.

Faux positif : Séquence non affectée détectée comme THG.

Vrai négatif : Séquence non affectée détectée comme non-THG.

Faux négatif : Séquence affectée détectée comme non-THG.

Spécificité : Vrais négatifs/(Vrais positifs + Faux négatifs)

Sensibilité : Vrais positifs/(Vrais négatifs + Faux positifs)

Perspectives

Pour être en mesure d'améliorer cette méthode de détection des séquences transférées, plusieurs méthodes pourraient être tentées. Entre autres, à la place d'enlever toutes les séquences qui dépassent le seuil en une seule étape, ceci pourrait se faire une séquence à

la fois, par exemple, d'enlever la séquence affectant le plus le coefficient en premier, refaire les mesures et enlever le second, ainsi de suite. Il pourrait également s'agir d'identifier certains groupes de séquences affectant particulièrement le coefficient de corrélation, d'enlever les séquences formant ce groupe, et refaire les mesures. Une autre solution pourrait s'agir de regrouper les séquences en groupes monophylétiques et de tester comment ces groupes affectent les coefficients.

Annexe 3 - Résultats reliés au protocole présenté à l'annexe 1

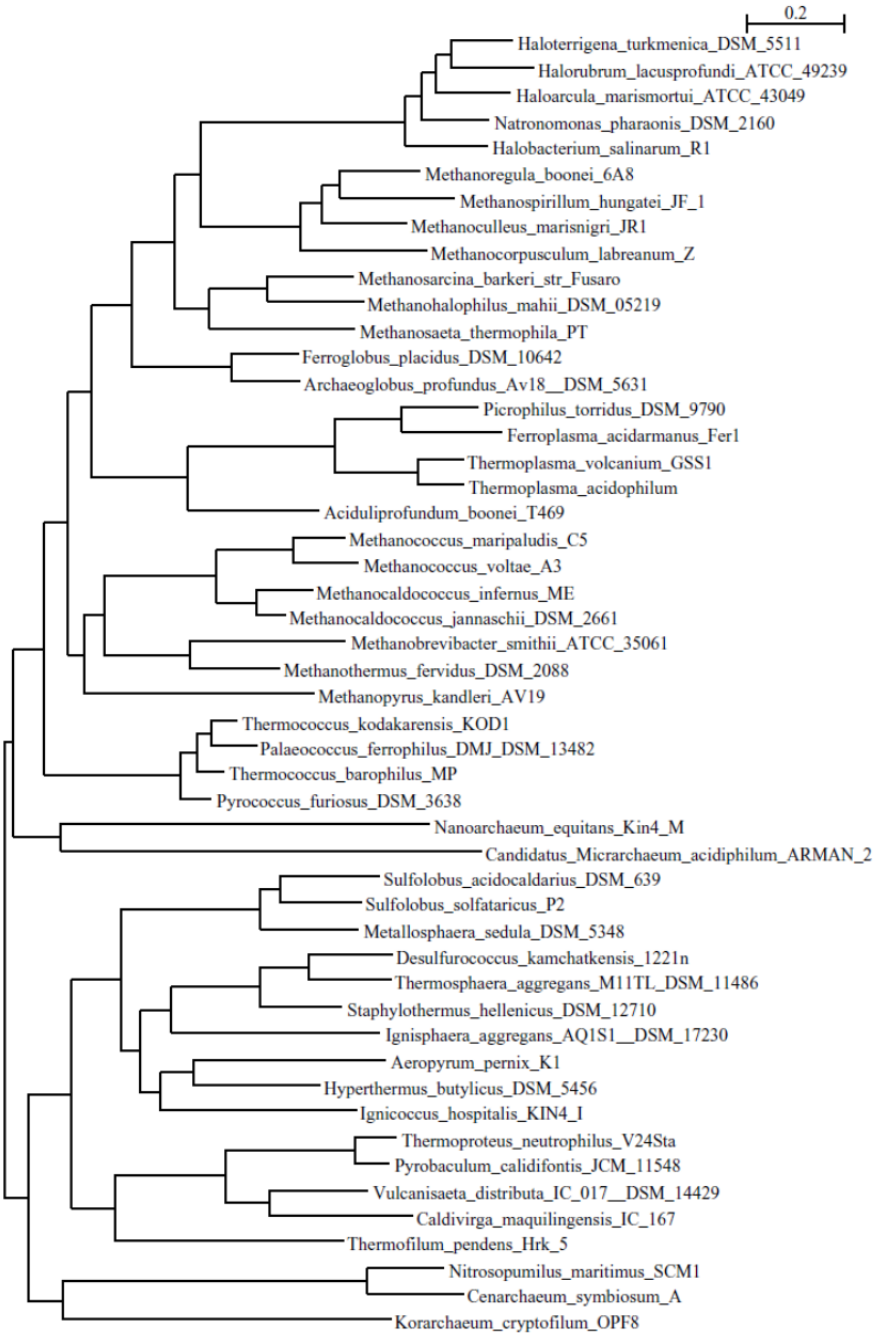


Figure A3.1 Arbre phylogénétique représentant l'histoire évolutive des Archées. Cet arbre est basé sur l'analyse 207 gènes et l'alignement concaténé contient 35349 positions. L'inférence a été effectué à l'aide de RAxML (Stamatakis, 2006) avec le modèle LG+Γ4 en estimant les fréquences empiriques des acides aminés.

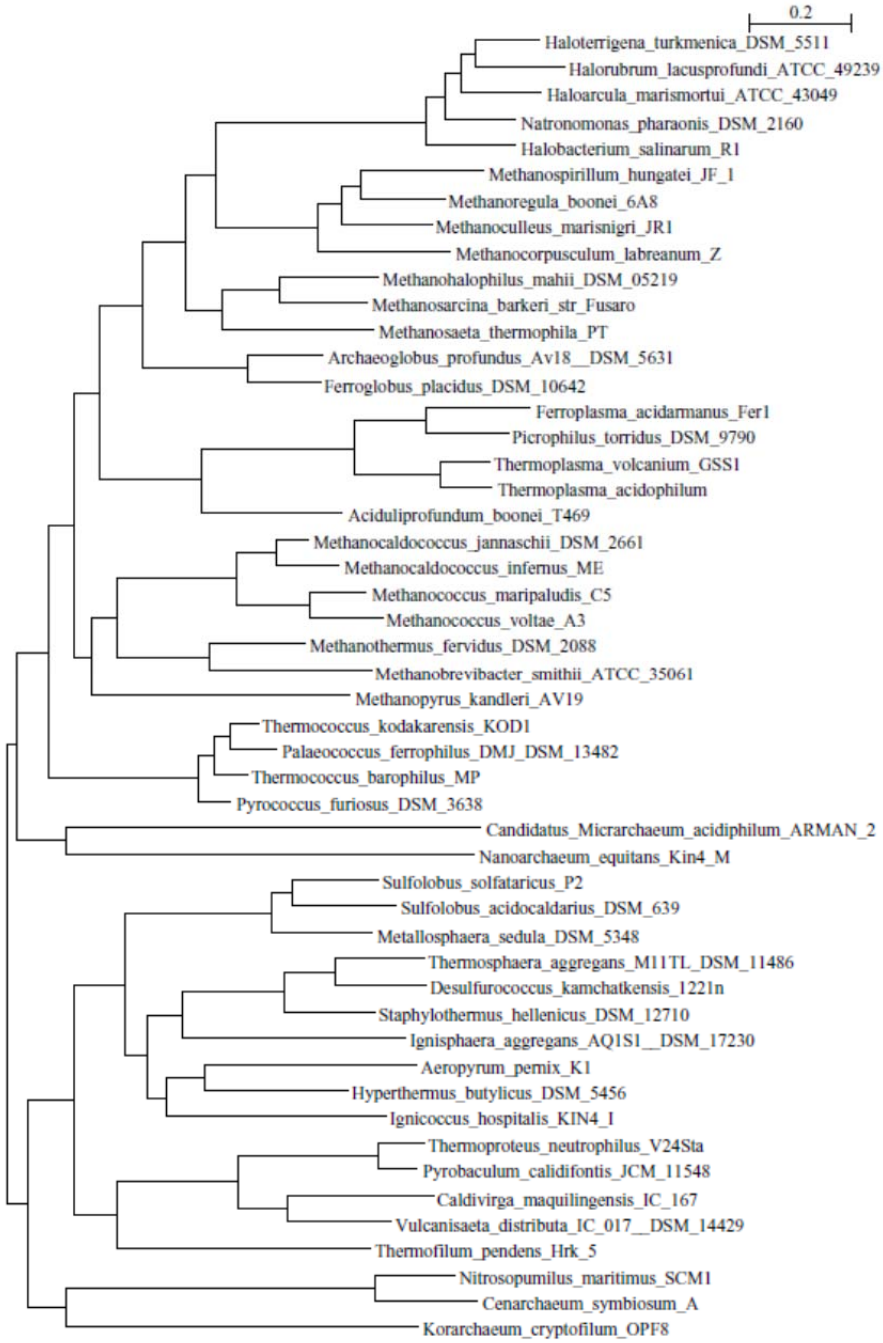


Figure A3.2 Arbre phylogénétique représentant l'histoire évolutive des Archées. Cet arbre est basé sur l'analyse 154 gènes et l'alignement concaténé contient 29218 positions. Il s'agit des mêmes gènes que présentés dans la figure annexe 3.1, sauf que les protéines ribosomiques ont été éliminées. L'inférence a été effectué à l'aide de RAxML (Stamatakis, 2006) avec le modèle LG+Γ4 en estimant les fréquences empiriques des acides aminés.

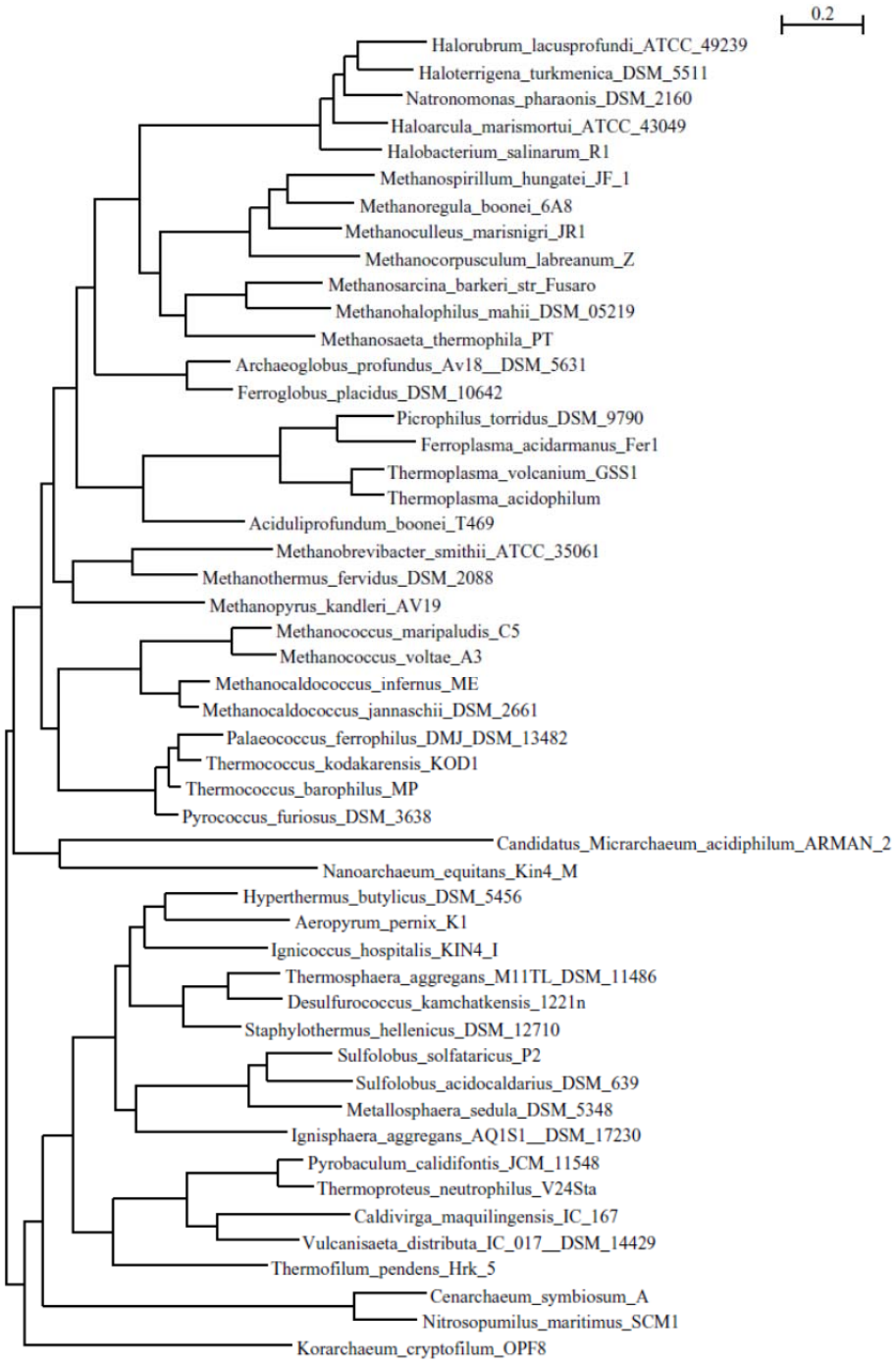


Figure A3.3 Arbre phylogénétique représentant l'histoire évolutive des Archées. Cet arbre est basé sur l'analyse 53 protéines ribosomiques et l'alignement concaténé contient 6131 positions. L'inférence a été effectué à l'aide de RAxML (Stamatakis, 2006) avec le modèle LG+Γ4 en estimant les fréquences empiriques des acides aminés.

Annexe 4 - Résultats reliés au protocole présenté à l'annexe 2

Les résultats obtenus ne sont pas encore très concluants (Table I). Bien sûr, un évènement de THG peut impliquer plusieurs espèces et peut être très difficile à détecter si l'évènement survient proche de la racine de l'arbre (Figure 1). La sensibilité, qui est le pourcentage de vrais transferts découvert, chute plus il y a d'espèces impliquées dans les évènements de transferts. La spécificité n'en est pas trop affectée.

Le suivi des évènements de THG par les simulations peut indiquer des processus de transferts multiples, qui sont très complexes et peuvent également revenir à leur position initiale, ce qui n'est donc pas possible de détecter. Mis à part les problèmes liés à l'écriture des évènements de simulations, la méthode de détection elle-même a encore des améliorations à être apportées.

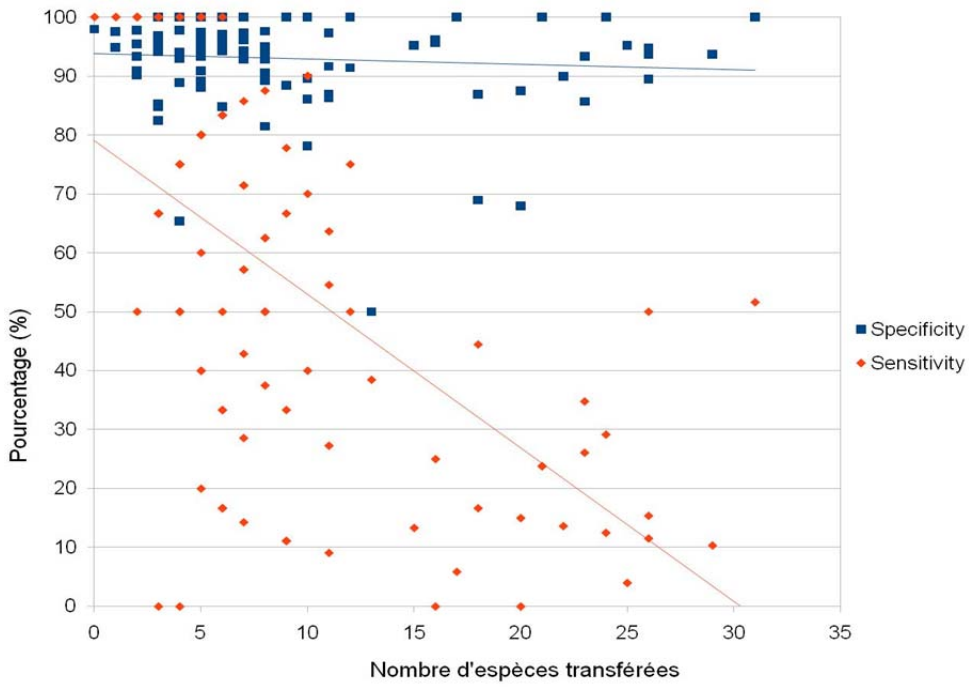


Figure A4.1 Évaluation de la spécificité et de la sensibilité selon le nombre d'espèces transférées pour 5 évènements de transferts horizontaux.

Table A4.1 Performance de détections de transferts horizontaux selon la méthode de filtration des alignements de séquences basée sur la comparaison de matrices de distances. Ces résultats sont basés sur 10 réplicats de 100 gènes concaténés.

	Specificity(%)	**Sensitivity (%)**
0 HGT	**92.1**	**n.a.**
5 HTGs	**89.8**	**50.72**
10 HGTs	**86.5**	**46.14**
20 HGTs	**76.8**	**41.34**
30 HGTs	**75.0**	**41.8**
40 HGTs	**53.6**	**41.8**
50 HGTs	**57.3**	**44.64**