Early Evolution and Phylogeny

Bastien Boussau

N° 205-2008 Année 2008 - 2009

THÈSE

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD - LYON 1

pour l'obtention

du DIPLÔME DE DOCTORAT

(arrêté du 7 août 2006)

soutenance prévue le
3 novembre 2008

par

Bastien BOUSSAU

Early Evolution and Phylogeny

Directeur de thèse: Manolo GOUY

Jury:	Laurent DURET	Examinateur
	Patrick FORTERRE	Examinateur
	Manolo GOUY	Directeur de thèse
	Didier PIAU	Rapporteur
	Ziheng YANG	Rapporteur

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J. F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur J. LIETO
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. SIMON
Secrétaire Général	**M. G. GAY**

SECTEUR SANTÉ

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur: M. le Professeur P. COCHAT
UFR de Médecine Lyon Grange-Blanche	Directeur: M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur: M. le Professeur J. ETIENNE
UFR de Médecine Lyon-Sud	Directeur: M. le Professeur F.N. GILLY
UFR d'Ontologie	Directeur: M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur: M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur: M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur: M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur: Mme. le Professeur S. FLECK
UFR de Biologie	Directeur: M. le Professeur H. PINON
UFR de Mécanique	Directeur: M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur: M. le Professeur G. CLERC
UFR de Sciences de la Terre	Directeur: M. le Professeur P. HANTZPERGUE
UFR de Mathématique	Directeur: M. le Professeur A. GOLDMAN
UFR d'Informatique	Directeur: M. le Professeur S. AKKOUCHE
UFR de Chimie Biochimie	Directeur: Mme. le Professeur H. PARROT
UFR STAPS	Directeur: M. le Professeur M.C. COLLIGNON
Observatoire de Lyon	Directeur: M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur: M. le Professeur J. LIETO
IUT A	Directeur: M. le Professeur M. C. COULET
IUT B	Directeur: M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur: M. le Professeur J. C. AUGROS

Remerciements

Je tiens tout d'abord à remercier Manolo Gouy pour son immense générosité, tout au long des quatre années où nous avons travaillé ensemble. Manolo m'a beaucoup donné : d'abord un excellent sujet de recherche, ensuite des idées brillantes et des conseils savants, une grande liberté, de nombreux encouragements. Je compte mener ma vie professionnelle et ma vie personnelle en le prenant comme modèle, et j'espère comme lui réussir une grande et belle carrière scientifique sans sacrifier honnêteté, rigueur morale, et vie familiale.

Il me faut également remercier Vincent Daubin. C'est en effet à son contact, et par contraste, que j'ai pu apprécier pleinement les qualités de Manolo. Lorsque j'ai accepté son invitation à partager son bureau, j'imaginais une atmosphère de travail saine et stimulante, des idées s'échangeant librement, des collaborations naturelles et harmonieuses. Vincent a au contraire toujours su agir en parasite, méprisant les idées qu'un des autres membres du bureau osait émettre, pour mieux les présenter marquées de son sceau lors d'une réunion ad'hoc dont il est encore aujourd'hui le chefaillon acariâtre et castrateur. Je crains de ne pas réussir à l'empêcher d'apposer son nom sur plusieurs de mes travaux à venir; même à Berkeley, j'aurai du mal à me défaire de son influence. (NB: Vincent était contre l'idée que je fasse des remerciements gentils, car ils seraient "trop classiques"; spécialement pour lui, j'ai donc copié les siens, mais je ne suis pas sûr d'y avoir beaucoup gagné en originalité.)

Je tiens aussi à remercier Anamaria Necşulea pour son aide et son enthousiasme tout au long de nos thèses. Les discussions avec Anouk sont toujours des plus revigorantes, car elle vous attribue toutes les idées dont elle est auteure, tous les progrès qu'elle a enclenchés. Il est pourtant évident pour tous ceux qui l'ont côtoyée qu'elle n'a besoin de personne pour mener des recherches d'une très grande qualité. C'est un plaisir de collaborer avec Anouk et je ne doute pas qu'elle atteindra, dans les mois à venir, les facteurs d'impact qu'elle mérite, à savoir les plus hauts.

C'est un plaisir de remercier Laurent Duret, dont la générosité intellectuelle autant que pécunière nourrit tout le laboratoire (et moi tout particulièrement ces derniers mois). J'ai beaucoup appris sur la science en discutant avec lui. Sa première moitié de carrière me paraît exemplaire, et j'ai hâte de voir la suite ! C'est un honneur qu'il ait accepté de présider mon jury de thèse.

Je voudrais également saluer les autres membres de mon jury de thèse, Patrick Forterre, Didier Piau et Ziheng Yang, pour avoir accepté de juger mon travail, et l'avoir fait avec beaucoup d'intelligence et de bienveillance. Leur apport a grandement amélioré le manuscrit. Qu'ils en soient remerciés.

Je remercie également Christian Gautier et Dominique Mouchiroud de m'avoir accueilli dans leur laboratoire, et d'y assurer une ambiance de travail cordiale et efficace. Je voudrais encore remercier Misou Pieri, Nathalie Arbasetti et Isabelle

Ravis pour m'avoir rendu la vie administrative douce et agréable.

Je remercie aussi Stéphane Delmotte, Lionel Humblot, Simon Penel, et Bruno Spataro pour leur grande tolérance concernant ma consommation déraisonnée de ressources informatiques, leur disponibilité de tous les instants et leur maîtrise sans faille des arcanes du binaire. Merci tout particulièrement à Simon pour sa représentation hyperréaliste de LUCA, cette image hante mes rêves.

Merci à tous les gens avec qui j'ai eu la chance de collaborer : Laurent Guéguen, sa segmentation et ses cours de Python, Céline Brochier-Armanet, Simonetta Gribaldo et Patrick Forterre qui m'ont très généreusement accueilli dans un de leurs projets (et merci à Céline pour son aide précieuse sur d'autres projets), Julien Dutheil, sa grande patience et ses formidables talents d'informaticien-biologiste-pédagogue, le brillant et néanmoins modeste Nicolas Lartillot, Samuel Blanquart et ses heures de calcul, Frédéric Brunet et Vincent Laudet et leurs poissons dupliqués, Siv Andersson et Marc Robinson-Rechavi qui m'ont mis le pied à l'étrier.

Je remercie encore Eric Tannier, un historien aux stupéfiantes connaissances en informatique et en biologie, Sophie Abby et son puits de mp3, Jean-François Gout, un gaucher qui coupe un peu trop au mien (de goût), Leonor Palmeira, qui une fois me surprit en chutant bruyamment mais dignement, Claire Guillet, Alexandra Popa, Joan Ho-Huu, Yves Clément et Thomas Bigot (les filles d'à côté), qui ont toléré mes fréquentes intrusions dans leur bureau à la porte fermée, Gabriel Marais et Raquel Tavares et leur généreuse invitation dans leur paradis nudiste, Sylvain Mousset et son ordinateur qu'il espère revoir un jour, Anne-Sophie Sertier pour sa bonne humeur et ses produits locaux auxquels je n'ose pas toucher, Vincent Lombard et son affection démonstrative, Anne-Muriel Arigon et Alexandra Calteau pour les bons moments qu'on a passés ensemble et qu'on passe ensemble lorsqu'on se retrouve en congrès, Jean Lobry pour ses conseils sur les thermomètres et oxygénomètres, Jean Thioulouse pour ses conseils en analyses statistiques, Guy Perrière pour m'avoir initié aux steaks de chez Chili's à Phoenix, Daniel Kahn et son insatiable curiosité scientifique, Marie Sémon et ses grenouilles, Philippe Veber et ses remerciements expéditifs.

Je tiens à remercier ma famille, qui s'est déplacée en grand nombre et m'a soutenu avec force objets magiques. Je suis particulièrement reconnaissant à mes parents d'avoir su accepter au sein de leur famille un scientifique, de toute évidence un monstre qui n'avait rien de prometteur. Merci à Emilie de m'avoir appris l'anglais avant l'heure avec les Incollables, cela s'est révélé utile finalement, et merci à David pour avoir lu avec moi les "Sciences-et-Vie Junior" : c'est aussi un peu à lui que je dois d'avoir fait sciences plutôt que lettres.

Enfin je remercie Mathilde, pour son tendre soutien, sa logique infaillible, et sa résistance au froid, et qui supporte mes crises de stress depuis tant d'années avec un dévouement qui force l'admiration. J'aimerais qu'elle continue.

Contents

1 Résumé en français

1.1 Tête de pont

Figure 1.1: *Les Demoiselles d'Avignon, peinture de Pablo Picasso, 1907, Museum of Modern Art, New York. Récupéré de Wikipedia, http://en.wikipedia.org/wiki/Image:Chicks-from-avignon.jpg*

La figure 1.1 a quelque-chose d'étonnant. En effet, dans cette peinture de Pablo Picasso, les deux demoiselles sur la droite ont des visages fort différents de leurs consoeurs. Ceux-ci apparaissent comme décomposés, aplatis. S'il est probablement illusoire de chercher à comprendre en détails pourquoi l'auteur a décidé de peindre ainsi ces demoiselles-ci en particulier, on peut toutefois essayer de savoir d'où lui est venue cette étrange inspiration.

Cette information peut être trouvée dans sa biographie : au début du vingtième siècle, plusieurs artistes, dont Gauguin, Matisse et Picasso, ont été profondément inspirés par la découverte d'arts anciens, ibérique et africain notamment. Ainsi, en 1905, une exposition consacrée à l'art ibérique fut montée au Louvre, et à l'automne 1906, Matisse montra une statuette africaine à Picasso, laquelle lui fit dit-on forte impression. Dans les traditions ibériques et africaines (il est abusif de parler d'art africain comme d'un tout homogène néanmoins, car on y trouve une grande diversité), les formes sont très stylisées, et l'importance d'un élément est souvent représentée par sa taille. Les deux visages des demoiselles de la droite, avec leurs grands yeux simples et démesurés ressemblent bien à des statues africaines ou ibériques. Dés lors, on comprend que ces visages aplatis et décomposés, qui préfigurent le mouvement cubiste, sont influencés par ces formes d'art. Mieux comprendre l'art de Picasso nécessite de se plonger dans son histoire.

1.2 L'histoire en sciences de la vie

Il n'y a pas qu'en arts plastiques que l'histoire permet de mieux comprendre une observation. Notre monde physique est soumis à l'empreinte du temps, et tous les phénomènes aujourd'hui observés sont le fruit d'une histoire, mêlant hasard et nécessité. Naturellement, le vivant ne fait pas exception, et c'est dans son histoire que l'on peut comprendre comment sont apparues les formes de vie retrouvées dans les fossiles ou observées de nos jours.

En biologie, la théorie de l'évolution permet d'expliquer la répartition et l'organisation des organismes vivants, ou plus précisément comment les êtres vivants ont acquis leur répartition géographique et écologique actuelle, et comment ils ont acquis leurs caractéristiques actuelles de forme, de fonction. Cette théorie est basée sur de nombreuses données que je ne détaillerai pas ici, mais peut être chichement résumée en quelques points :

- Tous les êtres vivants (plantes, Bactéries, homme, archées productrices de méthane dans l'estomac des vaches...) sont apparentés. En effet, quand on regarde comment fonctionnent tous ces organismes en détails, au niveau moléculaire, on se rend compte que tous les êtres vivants sont très ressemblants, ce qui trahit leur ascendance commune. De même, tous les organismes vivants sont construits autour d'un génome fait d'ADN, qui contient tous les gènes d'un organisme, et qui renferme toutes les recettes de cuisine nécessaires à la construction et au fonctionnement d'un être vivant. Or, quand on compare les génomes des différents êtres vivants entre eux (et c'est entre autres l'objet de cette thèse), on se rend compte là-encore qu'il

y a de grandes similarités entre tous les êtres vivants. Corollaire : tous les êtres vivants descendent d'un lointain ancêtre commun, que l'on appelle souvent $LUCA$, ce qui correspond à Last Universal Common Ancestor, soit dernier ancêtre commun universel.

- Les êtres vivants ont deux types de caractéristiques, des caractéristiques innées qui émanent directement de leur génome, et des caractéristiques acquises, qui sont le fruit de leur histoire personnelle. Seules les caractéristiques innées peuvent être transmises à leur descendance au travers des mécanismes de l'*hérédité*.

- Au cours de la transmission de ces caractéristiques innées, des *mutations* et des réarrangements peuvent survenir, ce qui a pour conséquence qu'un descendant, même s'il est très semblable, est très rarement rigoureusement identique à son (ou ses) géniteur(s), parce que leurs génomes diffèrent. Avec le temps, les générations succédant aux générations, les mutations s'accumulent, les génomes ressemblent de moins en moins au génome de l'ancêtre commun, et par conséquent les descendants ressemblent de moins en moins à leur lointain aïeul.

- Ces changements au cours de la transmission des caractères innés font qu'aucun être vivant n'est la copie parfaite d'un autre. On dit alors qu'il y a une grande *diversité* entre êtres vivants. Cette diversité fait que, étant donné un environnement, certains êtres vivants ont plus de facilités à avoir des descendants, et donc peuvent en avoir plus, que d'autres. Ces facilités sont généralement regroupées sous le terme anglais de fitness, qui correspond à la capacité *a priori* d'un organisme à se reproduire, son adaptation à son environnement. A un instant donné, les organismes ayant une plus grande fitness ont en moyenne plus de descendants, mécanisme que l'on nomme en général *sélection darwinienne*.

- Les organismes les plus adaptés n'ont pourtant pas toujours plus de descendants que les autres : on s'attend à ce qu'ils en aient plus, mais si par malheur la foudre s'abat sur eux avant qu'ils n'atteignent leur âge adulte, leur fitness aussi grande soit-elle n'aura pas eu beaucoup d'effet. Il y a donc un effet important du hasard sur qui se reproduit et qui ne se reproduit pas dans une population d'êtres vivants. Cet effet du hasard est d'autant plus fort que le nombre d'individus dans la population est faible : si on a 10% d'individus très adaptés dans une population au total de seulement 10 individus, il suffit d'un éclair pour que le meilleur d'entre eux n'ait pas de descendant. On appelle cet effet aléatoire la *dérive génétique*, puisqu'il fait dériver la fitness moyenne d'une population loin de ce qu'elle aurait été sans lui.

L'évolution des êtres vivants est donc un mélange de plusieurs mécanismes. Deux sont aléatoires : les mutations d'une part, et la dérive (sur qui tombe la foudre) de l'autre. Le troisième est déterministe, et fait que certains individus ont à la naissance, au regard de leurs caractéristiques génétiques et de l'environnement présent, plus de chances d'avoir des descendants que d'autres. Ces trois mécanismes s'associent et produisent la diversité entre individus et la diversité entre espèces que l'on peut observer aujourd'hui.

1.3 Mécanisme et phylogénie

On peut étudier deux aspects de l'évolution :

1. D'une part, on peut s'intéresser aux changements qui sont survenus au cours de l'histoire de la vie, et leur cause, soit (mutation+sélection), ou bien (mutation+dérive). J'appellerai cet aspect le *mécanisme* de l'évolution.

2. D'autre part, on peut chercher à décrire les relations de parenté entre êtres vivants. Cette représentation des relations de parenté entre organismes s'appelle la *phylogénie*.

Au cours de ma thèse je me suis intéressé à ces deux aspects de l'évolution. J'ai cherché à préciser certaines relations de parenté, et je me suis également attaché à découvrir certains changements qui ont pu se produire dans le passé. En fait, il est naturel de s'intéresser aux deux à la fois, car ils sont très dépendants. En effet, on ne s'intéresse aux mécanismes de l'évolution que dans le cadre d'une phylogénie particulière : si l'on plaçait les chauves-souris parmi les oiseaux et non parmi les mammifères (problème qui relève de la phylogénie), on ne se demanderait pas par quel mécanisme elles ont acquis leurs ailes, mais plutôt par quel mécanisme elles ont acquis leurs mammelles.

1.4 Une brève histoire de la vie

Les géologues estiment que la terre a plus de 4.5 milliards d'années, et que la vie y existe depuis au moins 3.5 milliards d'années (Schopf, 2006), peut-être plus. L'arrière arrière arrière ... grand père de tous les êtres aujourd'hui vivants, LUCA, date donc probablement de cette époque. Les descendants de LUCA ont ensuite donné naissance à d'autres organismes et, le temps aidant et les mutations s'accumulant, à de nouvelles espèces, possédant des caractéristiques inédites. 3.5 milliards d'années plus tard, tous les êtres vivants sont les descendants de ces premiers organismes.

1.4.1 L'analyse des génomes permet de reconstruire l'histoire de la vie

Afin de représenter cette généalogie universelle, l'arbre de la vie, qui représente les relations de parenté entre tous les êtres vivants, on peut analyser les ressemblances et différences entre les formes des êtres vivants, de la même façon que l'on pourrait essayer de reconstituer un arbre généalogique en analysant les différences physiques entre frères, soeurs, oncles, tantes et grands parents. Néanmoins cette approche n'est pas très aisée, surtout lorsqu'on cherche à comparer des plantes avec des animaux, des champignons, des bactéries... Depuis la fin des années 1970, d'immenses progrès ont été faits dans le séquençage de l'ADN, et l'on peut désormais séquencer des génomes entiers. On peut ainsi analyser les génomes des êtres vivants et les comparer, afin de reconstruire les relations de parenté. Cette dernière approche s'avère bien plus pratique.

Le génome d'un organisme contient toutes les recettes de cuisine utiles pour produire et faire fonctionner cet organisme. En comparant les génomes, on a donc directement accès à l'essence des caractères innés d'un organisme. Comme seuls les caractères innés sont transmis par l'hérédité, en analysant les génomes, on a accès à toute l'information qui a été transmise depuis LUCA jusqu'aux organismes actuels. Ces génomes portent les traces d'événements de mutation, sélection et dérive qui ont façonné les organismes vivants au cours de leur histoire, et constituent des documents de l'histoire évolutive d'une qualité unique: simplement en lisant des génomes, on peut tirer des conclusions sur les caractéristiques et l'histoire des organismes qui les contiennent. Encore faut-il savoir les lire.

Afin de lire ces documents, il faut faire un peu de mathématiques, en l'occurence des statistiques. A l'aide de modèles statistiques, il est possible d'estimer par exemple la probabilité que les champignons soient plus proches parents des animaux que des plantes, la probabilité qu'une mutation particulière se soit produite à un moment particulier dans l'arbre de la vie, la probabilité que cette mutation ait été transmise à des descendants par sélection ou bien par dérive. On peut donc poser des questions qui relèvent de la phylogénie ou bien du mécanisme de l'évolution.

1.4.2 L'arbre de la vie comme raconté par les génomes

L'analyse statistique des génomes a permis de découvrir que les êtres vivants se rangeaient dans trois grandes catégories (Woese et Fox, 1977) : les Archées, les Bactéries, et les Eukaryotes.

- Les Archées comprennent des organismes composés d'une seule cellule, que l'on trouve un peu partout mais aussi dans des milieux très insolites,

depuis la panse des vaches où elles aident à la digestion en dégageant du méthane, jusqu'aux sources thermales où des archées peuvent vivre à plus de 100°C, en passant par des milieux extrêmement acides ou bien saturés en sel (Forterre, 2007).

- Les Bactéries elles aussi sont souvent composées d'une seule cellule, vivent dans toutes sortes de milieux, mais elles sont en général moins exubérantes dans leurs préférences écologiques. On trouve notamment parmi les Bactéries les Cyanobactéries, qui peuvent utiliser l'énergie lumineuse pour vivre et qui produisent de l'oxygène. On trouve également parmi les Bactéries des parasites des plantes ou bien des animaux, certaines d'entre elles causant plusieurs maladies bien connues, comme le typhus, le cholera, la lèpre. Toutes les Bactéries ne sont toutefois pas parasitiques, et bon nombre d'entre elles sont nos symbiotes, nous aidant notamment à digérer.

- Enfin les Eucaryotes contiennent les êtres vivants les plus connus, et notamment les plus grands, qui peuvent posséder des milliards de cellules. Ils contiennent champignons, animaux, plantes, amibes, infusoires. Ils sont en général assez peu surprenants dans leurs goûts écologiques, n'appréciant guère les températures, PH ou salinités extrêmes. En outre, ils ne montrent pas beaucoup de types de métabolismes différents, puisqu'il n'y a en gros que deux types d'Eucaryotes de ce point de vue, ceux qui se reposent sur la photosynthèse, et ceux qui consomment la matière organique produite par d'autres êtres vivants. Les Eucaryotes ont en général des organites dans leurs cellules, petites structures qui renferment un génome particulier. Il existe notamment deux types d'organites : les mitochondries, où les sucres sont dégradés pour faire de l'énergie, et les chloroplastes chez les plantes, où les rayons lumineux sont transformés en énergie puis en sucres. La présence de plusieurs génomes dans une seule cellule a une explication historique que j'expliquerai un peu plus tard.

On pense actuellement que les Archées et les Eucaryotes sont de plus proches parents l'un de l'autre qu'ils ne le sont des Bactéries (Gogarten *et al.*, 1989; Iwabe *et al.*, 1989). De nombreux travaux suggèrent que l'arbre de la vie pourrait ressembler, au moins dans ses grandes lignes, à ce qui est représenté Fig. 1.2.

Les deux grandes flêches colorées qui traversent l'arbre en Fig.1.2 permettent d'expliquer la présence des organites chez les Eucaryotes. En analysant les génomes des mitochondries et chloroplastes, les phylogénéticiens ont pu montrer que ces organites étaient en fait d'anciennes Bactéries, kidnappées par les Eucaryotes (Zablen *et al.*, 1975; Bonen et Doolittle, 1975; Bonen *et al.*, 1977; Esser *et al.*, 2004; Deusch *et al.*, 2008) et réduites à l'esclavage. Au final, les Eucaryotes n'ont pas été imaginatifs en ce qui concerne leurs métabolismes, et en plus ils ont volé le peu qu'ils savent faire des Bactéries. On peut concevoir une hypothèse historique pour expliquer ce manque d'originalité : si les Bactéries étaient présentes

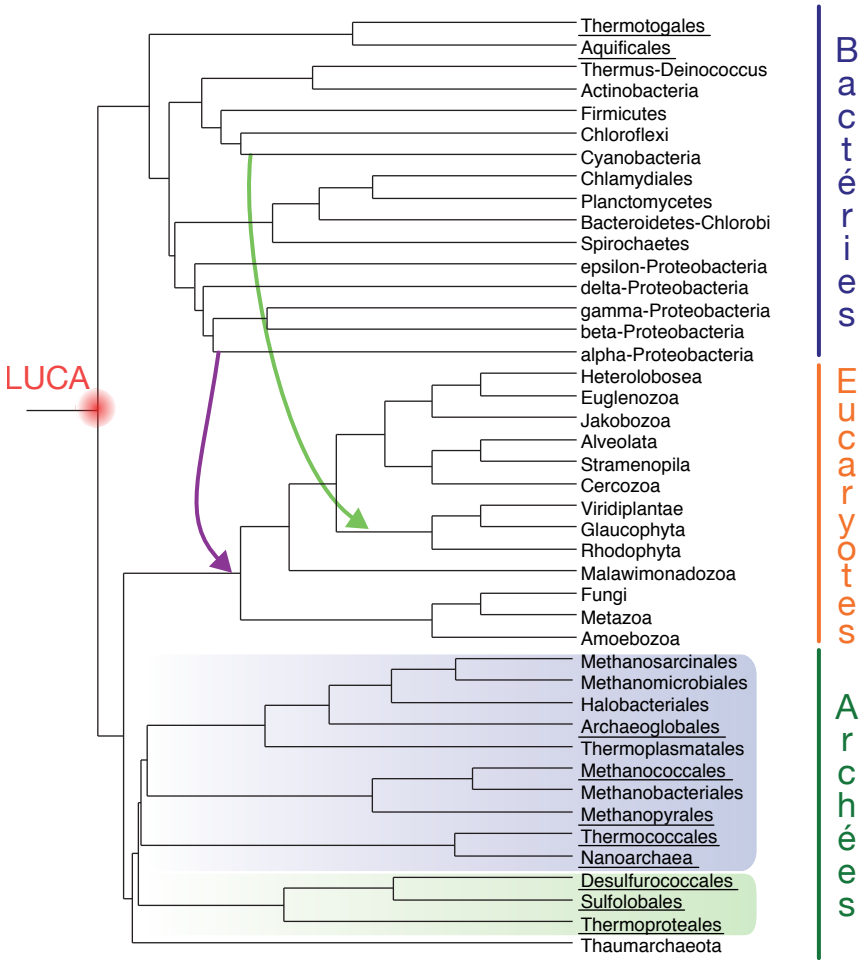


Figure 1.2: *Arbre de la vie. L'extrême gauche de l'arbre a un âge de plus de 3.5 milliards d'années, la partie droite concerne des organismes actuels. LUCA est représenté avec un point rouge. Les organismes dont le nom est souligné vivent à plus de* 80°C. *Les organismes figurés sur fond vert sont les Crenotes, et sur fond bleu les Euryotes, les deux grandes classes d'Archées. Tous les animaux sont classés au sein des Metazoa, tous les champignons se trouvent au sein des Fungi, et toutes les plantes sont placées dans les Viridiplantae. Tous les organismes vivants que l'on voit à l'oeil nu représentent donc une infime portion de la biodiversité.*

sur terre avant les Eucaryotes, alors la plupart des niches écologiques devaient déjà être occupées. Les Eucaryotes se sont donc spécialisés dans une autre sorte

d'activité : le vol et la prédation.

1.5 Mon travail de thèse

Au cours de ma thèse, je me suis intéressé à quelques problèmes particuliers ayant trait à l'étude des génomes pour reconstruire leur histoire. J'ai cherché à améliorer les méthodes de reconstruction de l'évolution des génomes, et j'ai utilisé ces méthodes pour répondre à des questions biologiques précises. Presque toutes ces questions sont liées à la façon dont un caractère particulier, la température préférée des organismes, a évolué.

J'ai signalé plus haut que certaines Archées pouvaient vivre à plus de 100°C ; en fait, la plupart des espèces ne sont capables de vivre que dans une petite fenêtre de températures : à 5 ou 10 degrés près, un organisme peut ne plus se développer normalement voire dépérir. Certains organismes ne peuvent donc vivre qu'aux alentours de 37°C, d'autres qu'autour de 10°C, de 100°C, *etc.* On caractérise généralement les organismes par leur température optimale de croissance, la température à laquelle leur croissance est la plus rapide. Cette température est un paramètre important : à une température donnée correspond un environnement particulier. Si un organisme a une très haute température optimale de croissance, on sait qu'il vit proche d'une source thermale comme dans le parc de Yellowstone ou bien comme au niveau des dorsales océaniques ; s'il vit à très faible température, on a également une idée assez précise des endroits où il pourrait vivre.

Dans la figure 1.2, les organismes dont le nom a été souligné ont des températures optimales de croissance supérieures à 80°C, ce qui fait qu'on les appelle des hyperthermophiles. Leur répartition dans l'arbre de la vie est intriguante : de nombreuses archées sont hyperthermophiles, ce qui indiquerait que l'ancêtre de toutes les archées était lui-même hyperthermophile. De même, deux bactéries situées à la base du domaine bactérien sont hyperthermophiles, ce qui suggère que l'ancêtre des Bactéries vivait peut-être à haute température. Si les ancêtres des Bactéries et des Archées étaient tous deux hyperthermophiles, alors notre grand père à tous, LUCA, lui aussi pourrait avoir vécu à très haute température, et la vie serait donc intimement liée à ces environnements extrêmes.

Je me suis donc attaché à étudier la phylogénie de ces deux grands domaines, les Archées et les Bactéries, pour étudier les positions d'organismes dont la température optimale de croissance est clé pour la reconstruction de l'évolution de ce caractère. Plus particulièrement, j'ai étudié la position de la Bactérie *Aquifex aeolicus* (groupe Aquificales sur la figure 1.2), car il n'était pas évident que sa po-

sition à proximité des Thermotogales ne soit pas erronée : l'évolution du génome de ces bactéries hyperthermophiles est en effet très compliquée. Bien que je n'ai pas réussi à pleinement surmonter toutes les difficultés associées à l'analyse du génome d'*Aquifex aeolicus*, mon travail confirme que les Aquificales pourraient être proches parentes des Thermotogales. Chez les Archées, sur l'invitation de Céline Brochier, Simonetta Gribaldo et Patrick Forterre, je me suis intéressé à la position de *Cenarchaeum symbiosum* (groupe Thaumarchaeota sur la figure 1.2), dont la température optimale de croissance est de 10°C. L'analyse suggère que cette archée est bien différente des autres, qu'elle ne peut pas être affectée simplement à un des deux grands groupes connus d'Archées, les Euryotes et les Crénotes, mais qu'elle constituerait peut-être un branchement très basal de l'arbre des Archées. Du fait de sa faible température optimale de croissance, ce branchement semble remettre en cause l'idée selon laquelle l'ancêtre des Archées était probablement hyperthermophile.

Afin d'étudier l'évolution des températures optimales de croissance, j'ai également suivi une approche plus directe. Cette température est une caractéristique commune à toute une espèce, et donc émane du génome des organismes ; il existe d'ailleurs des moyens de prédire avec un petit peu de statistiques et un ordinateur, simplement à partir de la séquence du génome d'un organisme, quelle est sa température optimale de croissance. Cela signifie que si on est capable de reconstruire les séquences de génomes, ou même simplement de morceaux de génomes, d'anciens organismes, on peut prédire à quelle température vivaient ces organismes. En collaboration avec des chercheurs de Montpellier, j'ai pu ainsi estimer l'évolution des températures de croissance optimale au cours des derniers 3.5 milliards d'années, en reconstruisant les séquences de morceaux de génomes ancestraux. Les résultats que nous avons obtenus sont représentés Fig. 1.3.

Nos résultats suggèrent que LUCA ne vivait pas à très haute température, mais que ses deux descendants les ancêtres des Archées et des Bactéries vivaient à plus haute température que lui. Ensuite, chez les Bactéries (au moins), les températures de croissance semblent avoir décru à nouveau. Cette décroissance a déjà été décrite chez les Bactéries en début d'année par Gaucher *et al.* (2008), qui l'ont interprétée comme étant corrélée à la température des océans au cours des 3.5 derniers milliards d'années. La température optimale de croissance des Bactéries aurait donc suivi la température moyenne à la surface de la terre. Les adaptations parallèles à de hautes températures depuis LUCA sont par contre nouvelles. Nous avons donc cherché des hypothèses pouvant expliquer ce phénomène. Certaines sont représentées Fig. 1.4.

Il est ainsi possible que LUCA ait vécu dans un environnement de température moyenne, et ait donné naissance à de nombreux organismes. Parmi ceux-ci, les mutations aidant, certains auraient été plus résistants aux fortes températures,

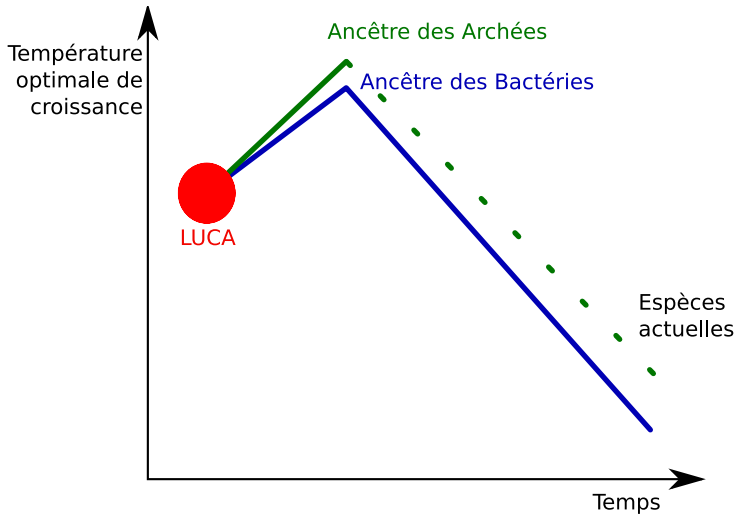


Figure 1.3: *Reconstruction des températures optimales de croissance le long de l'arbre de la vie. La deuxième partie de la courbe des Archées (vert) est en pointillé, cas nous n'avons pas assez de données pour la connaître avec suffisamment de certitude.*

d'autres moins. On sait qu'il y a 3.8 milliards d'années, la fréquence de chutes météoritiques a connu une grande augmentation (Gomes *et al.*, 2005) (épisode dit du Late Heavy Bombardment). Ces chutes de météorites ont probablement causé d'importants dégats sur terre, et ont du considérablement augmenter la température qui régnait à sa surface. Si LUCA avait vécu avant 3.8 milliards d'années, alors seuls les plus résistants à la chaleur de ses descendants auraient pu survivre. Ensuite, ces descendants auraient donné naissance aux Archées et Eucaryotes d'une part, et aux Bactéries d'autre part. Selon cette hypothèse, une pression de sélection liée à des chutes météoritiques serait à l'origine de l'évolution profonde des températures optimales de croissance.

Une autre hypothèse a été proposée par Forterre (2002) et suggère qu'une mutation aurait pu faciliter les adaptations à de plus grandes températures chez les descendants de LUCA. Selon cette hypothèse, LUCA avait un génome dont la molécule principale était l'ARN, et était donc sensible à la chaleur, alors que ses deux descendants ont chacun indépendamment acquis la possibilité d'utiliser l'ADN comme support de leur génome. Comme un génome à ADN serait plus résistant à la température qu'un génome à ARN, cette mutation aurait permis aux descendants de LUCA de vivre à de plus hautes températures.

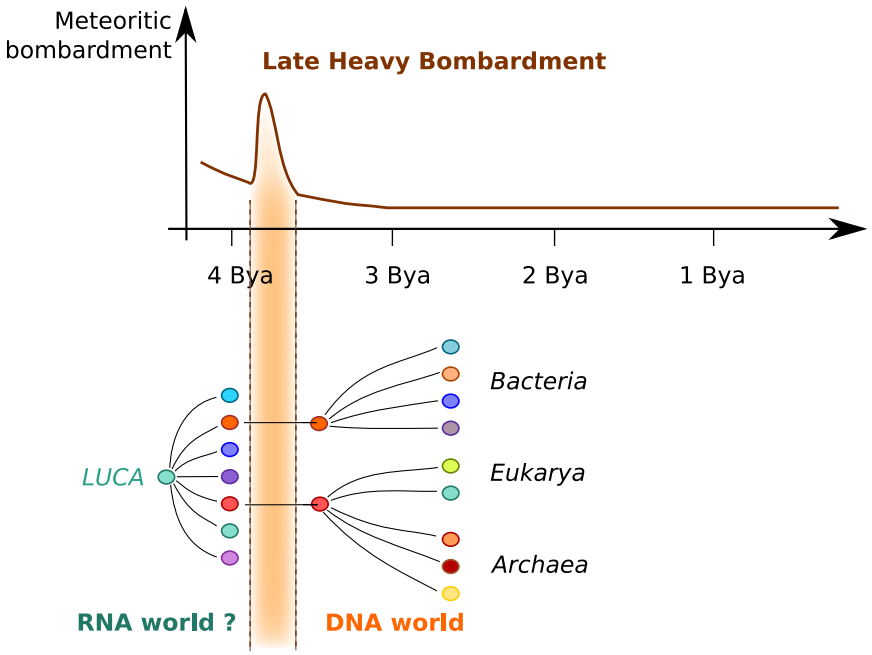


Figure 1.4: *Scénario pour l'évolution des températures optimales de croissance depuis LUCA jusqu'à ses descendants.*

1.6 Conclusion

Mon travail de thèse constitue un exemple de l'application de méthodes statistiques à l'analyse de génomes afin d'étudier l'évolution des organismes vivants. C'est une méthode puissante qui permet de traiter des questions qui sont inaccessibles à la plupart des autres sciences biologiques : la paléontologie notamment est limitée par la rareté, la petitesse, et la dégradation des fossiles.

Il y a toutefois de nombreuses difficultés associées à ces études, et ma thèse m'a convaincu qu'il fallait développer de meilleurs modèles statistiques d'évolution des génomes. En échange, il y aura beaucoup à apprendre sur l'évolution des génomes, des êtres vivants et de la terre.

2
Introduction

2.1 Time counts

There is a huge diversity of spoken languages in South America, some of which are endemic to this continent, and others also found elsewhere. If one considers only European idioms, Portuguese is spoken in Brazil, English in Guyana, French in French Guiana and Dutch in Suriname, while Spanish is the official language in most other countries. The presence and geographical distribution of European languages in a place far remote from Europe of course makes sense in the light of history: European languages have been brought to south America by European colonizers.

All things are affected by time.

In fact, all phenomena from the physical world obviously are the product of processes that unfolded through time: time is a major variable in our physical world, all objects are submitted to its footprint. Consequently, to explain the existence and organisation of natural objects, time must be accounted for.

Notably, as for languages, the diversity and distribution of species around the world can only be explained by looking at their history. In a famous passage of "The voyage of the Beagle" (Darwin, 1845), Charles Darwin describes finches of the *Geospiza* genus he sampled on the Galapagos islands, about 1,000 km West from South America:

> The most curious fact is the perfect gradation in the size of the beaks in the different species of Geospiza, from one as large as that of a hawfinch to that of a chaffinch, and (if Mr. Gould is right in including his sub-group, Certhidea, in the main group), even to that of a warbler.

He later turns towards a historical hypothesis to explain this diversity:

> Seeing this gradation and diversity of structure in one small, intimately related group of birds, one might really fancy that from an original paucity of birds in this archipelago, one species had been taken and modified for different ends.

Insights such as this one paved the way for the Origin of Species, which was published 20 years later. Darwin considered that the range of species observed on the Galapagos islands, with their similarities and geographical distribution, could not be explained by instantaneous independent creations followed by stasis, but was the trace of a historical process, which is now known as the theory of evolution (Darwin, 1859):

> The most striking and important fact for us in regard to the inhabitants of islands, is their affinity to those of the nearest mainland, without being actually the same species. Numerous instances could be given of this fact. I will give only one, that of the Galapagos Archipelago, situated under the equator, between 500 and 600 miles from the shores of South America. Here almost every product of the land and water bears the unmistakable stamp of the American continent. There are twenty-six land birds, and twenty-five of these are ranked by Mr. Gould as distinct species, supposed to have been created here; yet the close affinity of most of these birds to American species in every character, in their habits, gestures, and tones of voice, was manifest. So it is with the other animals, and with nearly all the plants, as shown by Dr. Hooker in his admirable memoir on the Flora of this archipelago. The naturalist, looking at the inhabitants of these volcanic islands in the Pacific, distant several hundred miles from the continent, yet feels that he is standing on American land. Why should this be so? why should the species which are supposed to have been created in the Galapagos Archipelago, and nowhere else, bear so plain a stamp of affinity to those created in America? There is nothing in the conditions of life, in the geological nature of the islands, in their height or climate, or in the proportions in which the several classes are associated together, which resembles closely the conditions of the South American coast: in fact there is a considerable dissimilarity in all these respects. On the other hand, there is a considerable degree of resemblance in the volcanic nature of the soil, in climate, height, and size of the islands, between the Galapagos and Cape de Verde Archipelagos: but what an entire and absolute difference in their inhabitants! The inhabitants of the Cape de Verde Islands are related to those of Africa, like those of the Galapagos to America. I believe this grand fact can receive no sort of explanation on the ordinary view of independent creation; whereas on the view here maintained, it is obvious that the Galapagos Islands would be likely to receive colonists, whether by occasional means of transport or by formerly continuous land, from America; and the Cape de Verde Islands from Africa; and that such colonists would be liable to modification;-the principle of inheritance still betraying their original birthplace.

It is now accepted that the organisms one observes are the product of a historical process. Considering that living matter has been shaped through time can explain many puzzling observations. For instance, some marine vertebrates have a horizontal caudal fin instead of a vertical caudal fin like most other ones, because one of their ancestor was a terrestrial tetrapod; the structure of the bones of our inner ear is partly explained by their origin as jaw bones in our fish- and reptile-like ancestors; stochasticity and convergent evolution explains why the mammalian fauna of Australia is largely dominated by marsupials, whereas other continents contain mainly placental mammals. This cliché quotation from Dobzhansky (1973) sums it up:

Life has been shaped by history

> Nothing in Biology Makes Sense Except in the Light of Evolution.

2.2 Pattern and process

Once it is accepted that evolution has shaped biological diversity, *i.e.* for instance the differences between species, the question remains as to how evolution did shape this diversity. There are two aspects to that question, that Galtier et Gouy (1998) named pattern and process. Pattern corresponds to the trajectories evolution has adopted to arrive at extant organisms, and is known as the phylogeny. A phylogeny depicts relationships between organisms through time, from a common ancestor to extant organisms; it shows which organisms are more closely related than others. Once the pattern is in place, process corresponds to how evolution did walk along these trajectories, *i.e.* what events occurred, and when.

Pattern corresponds to the history of speciations.

Process corresponds to the mechanistic of evolution along the tree.

2.2.1 Pattern

The phylogeny represents family relationships between species: it can be seen as a family tree of species (Fig. 2.1).

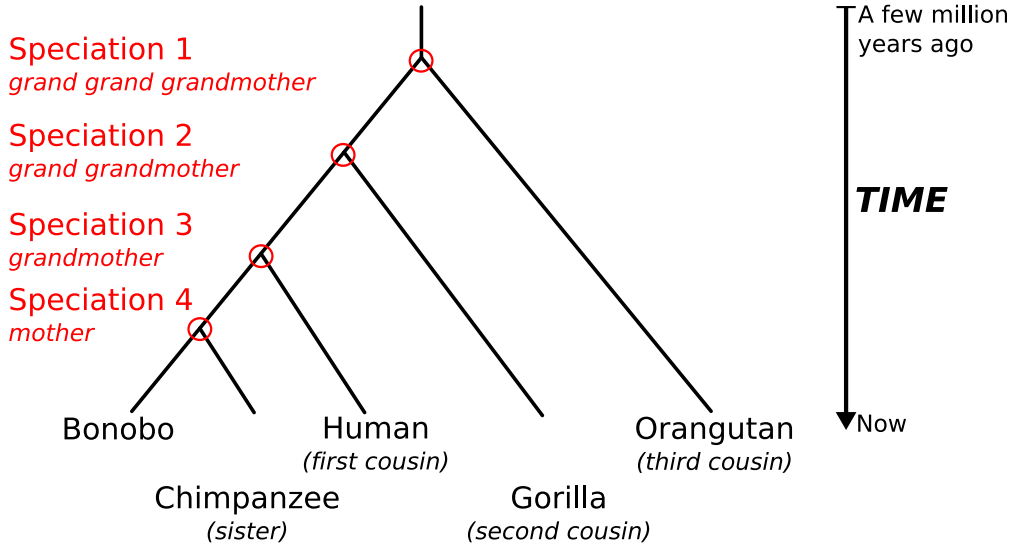


Figure 2.1: *A phylogeny of some apes. A phylogeny is very much like a family tree, where family members are species. Here the family tree is centred on the Bonobo chimpanzee. Ancestral nodes in a phylogenetic tree as well as in a family tree correspond to ancestors. In a phylogenetic tree, they also correspond to speciations. Here, the first speciation separates Orangutan from Gorilla, Humans and the two species of Chimpanzees.*

2.2.2 Process

Once the pattern is in place, one can use it as a framework for asking questions about the process of evolution, *i.e.* events that occurred along the routes evolution has taken, and that explain the similarities and differences among species. Questions related to the process of evolution could be: Did evolution go fast? Did evolution proceed towards increases in size? Was evolution totally stochastic so that no trend seems to be identifiable? In fact, both pattern and process are intimately related, and usually knowing one permits to improve the study of the other. Thus, when studying evolution, it is better to estimate at the same time pattern and process.

During my thesis, I have developed methods to reconstruct phylogenies and infer events along the branches of a phylogeny, and made efforts towards improving the tree of life, by studying particular species whose phylogenetic relationships are disputed. I have focused on more ancient events than the speciations of great apes, and notably on how the major classes of extant organisms came to be the way they are. In the next section, I explain why biologists think that all living beings share a common evolutionary history, and what are the major classes, or

kingdoms, of life.

2.3 LUCA and the three kingdoms

2.3.1 The unity of life

In Fig. 2.1, phylogenetic relationships between only 5 species of apes are represented; much larger phylogenies can be built, that encompass birds, fishes, insects, mollusks, mushrooms, plants, and all sorts of microorganisms. If such a large phylogeny does not forget any sort of living being, it corresponds to a *tree of life*. On Earth, all organisms can be included in such a tree of life, because they share a few characteristics that hint at their common origin: they have so many common points that it is more reasonable to assume that they inherited these properties from a common ancestor rather than all evolve independently the same characteristics. This means that all living beings that we now observe, grass, human, bacteria, heat-loving micro-organism, have the same grand grand grand... grandmother. This grandmother of all has been named *LUCA* for Last Universal Common Ancestor, and is the object of much interest and many controversies.

All life comes from LUCA.

Characters that show that all living beings have a single origin include their common basic organisational unit, the cell. Most Bacteria and other unicellular organisms are made of only one cell; multicellular organisms such as yourself may contain billions of them, that exchange information and interact to make the whole organism function. In both cases however, a cell is delimited by a lipidic membrane, that draws a boundary between the extracellular and the intracellular environments, filled with an intricate molecular menagerie. This menagerie includes all sorts of molecules, some used as energy currency or storage, others as infrastructures, others as nanoscopic machines, others as information storage... Then, the way that these molecules are used is also universal: general processes indispensable and central to the living of the cell are nearly identical in all organisms. Eventually, all these similarities come from large similarities in the genomes of all living beings, as genomes constitute an organism's cookbook. Such large similarities are best explained by the hypothesis of a common unique origin for life on Earth. Therefore, all organisms are related, and their relationships are represented by the tree of life.

Cellular and molecular structures are similar in all organisms.

Once it is understood that all life has a single origin, a giant family tree can be built, that encompasses all kinds of living beings. Although the endeavour to build such a tree is still very much a work in progress (AToL, 2008), some hypotheses of the basic organisation of the tree of life now seem to be nearly certain.

2.3.2 The three kingdoms

The primary divisions of life are Archaea, Bacteria and Eukarya.

One important analysis led Woese et Fox (1977) to propose that the major divisions of life, the "primary kingdoms", were Eubacteria (hereafter named Bacteria), Archaebacteria (Archaea), and Urkaryotes (Eukarya). This trichotomy has since then been mostly confirmed by more sophisticated analyses (see section 2.7 for more details).

Eukarya have a nucleus and organelles.

- Among the three kingdoms, Eukarya are the most conspicuous, as they contain most multicellular organisms. However, besides plants, animals, fungi and amoeba, Eukarya also contain many groups of unicellular species. All Eukarya have their DNA packed in a nucleus, and have their cells decorated with many organelles. These organelles are lipid-membrane individualised compartments that drive particular functions in the cell: for instance mitochondria house the process of respiration, *i.e.* the oxidation of particular molecules to yield Adenosin TriPhosphate (ATP), a small molecule that can then be used as a provider of energy for all sorts of reactions in the cell. Another well known example is the chloroplast, found in plants, and where photosynthesis takes place: photosynthesis transforms light energy into chemical energy, notably once again in the form of ATP. This chemical energy can then be used in the chloroplast to produce small sugars, that can be used or stored by the cell. Eukaryotic cells are further equipped with a sophisticated cytoskeleton, which, associated with the fact that many Eukaryotes do not have a cell wall, gives them the ability to change their shape and is a fundamental element of their ability to move or to engulf particles. It is also used during cell division or intracellular trafficking.

Many Archaea live in extreme environments.

- Archaea contain mainly unicellular species, of smaller size than eukaryotic cells, and have neither nucleus nor organelles. Most species are protected by a cell wall, and the composition of their membrane is different from that found in Eukarya and Bacteria. Contrary to Eukarya, Archaea display a wide range of metabolisms. Notably, methanogenesis that reduces CO_2 with H_2 to produce methane is only encountered among Archaea. Therefore, if cows produce such huge amounts of methane, it is thanks to the methanogenic archaea in their guts. Other Archaea are famous for their ability to cope with inhospitable environments such as those encountered in very hot volcanic sources, very salty ponds, or very acid mine wastes. Interestingly, although some symbiotic Archaea have been described (Preston *et al.*, 1996), no parasitic Archaea has been found so far. Recently, it has become clear that Archaea could play important roles in the economy of the Earth, by being major contributors to several biochemical cycles (Schimel,

2004; Leininger *et al.*, 2006; Lipp *et al.*, 2008).

- Bacteria cannot easily be differentiated from Archaea simply based on their morphology: accordingly, the three kingdoms were defined based on the analysis of a gene sequence, not cell structure. Indeed, they are similarly sized as Archaea, can be similarly shaped and do not contain organelles either. Bacterial cells are also protected by a cell wall, and also harbour a great diversity in their metabolisms and in their lifestyles. For instance, they invented oxygenic chlorophyll-or bacteriochlorophyll-based photosynthesis. Other types of metabolisms can also be found, with heterotrophic or autotrophic species. Both symbiotic and parasitic Bacteria have been discovered, with some pathogenic Bacteria very well known for the sinister diseases they cause, such as *Mycobacterium leprae* (leprosis), *Vibrio cholerae* (cholera), *Treponema pallidum* (Syphilis), *Bacillus anthracis* (anthrax), *Yersinia pestis* (bubonic plague).

Bacteria invented photosynthesis.

All these types of organisms are the product of a history that can be inferred through geology and phylogenetic analyses. Geology studies rocks that carry the stigmas of ancient events, and thus can provide clues concerning the environment and the living beings that reigned billions of years ago; such information is usually very partial and eroded but cannot be obtained by other means. In the next sections, I will present results obtained first through geology and second through phylogenetics, a science that nicely complements geology.

2.4 A short history of life on Earth, as told by rocks

The solar system may be older than 4.5675 billion years (the number of decimals is due to Connelly *et al.* (2008)). Ancient rocks suggest that life has existed on Earth for more than 3.5 billion years (Schopf, 2006) . This estimate is based on the analysis and the datation of stromatolites ("accretionary sedimentary structures, commonly thinly layered, megascopic and calcareous, interpreted to have been produced by the activities of mat-building communities of mucilage-secreting micro-organisms" (Schopf, 2006)), of microfossils (fossils of structures resembling cells), of molecular biomarkers (molecules interpreted as being diagnostic of a particular group of organisms) and of isotopic data (measure of the frequencies of various isotopes of a given atom; these frequencies can be affected by biological processes). I briefly describe these three types of methods and present some of the insights into the deep history of life they provided. These insights have been selected arbitrarily, and focus mainly on ancient history (older than 1 billion

Life is more than 3.5 billion years old.

years ago). Interested readers are invited to read the books by Knoll (2004) and Lane (2004) for more details.

2.4.1 Microfossils

Establishing the biological origins of stromatolites or microfossils is usually difficult, and can only be achieved through comparisons with more recent, uncontroversial examples. For instance, stromatolites can be convincingly described as coming from biological processes if they display an important diversity in their shape. Accordingly, Allwood *et al.* (2006) described seven different morphotypes among 3.43 billion year old stromatolites from Australia, arguing that it seems unlikely that non-biological processes would show such a diversity. The oldest stromatolites discovered so far may thus be 3.43 billion years old. Similarly, microfossils showing two micrometer-sized spheres next to each other are usually interpreted as an example of cell division. However, most convincing fossils are those that combine microscopic and macroscopic clues of biological origins, such as microfossils inside rocks resembling stromatolites. Such combined evidence can also be strengthened by molecular biomarkers.

Some stromatolites are 3.43 billion years old.

2.4.2 Molecular biomarkers

Living organisms produce molecules that cannot be obtained by non-biological processes, that are characteristic of their metabolism and are thus named biomarkers. For instance, Cyanobacteria use 2-methylbacteriohopanepolyols in their membrane; because derivatives of these molecules, named 2-methylhopanoids, have been found in sediments 2.7 billion years old, this suggests that Cyanobacteria already existed at the time ((Summons *et al.*, 1999; Brocks *et al.*, 1999; Summons *et al.*, 2006) . Similarly, because some sterols diagnostic of Eukaryotic membranes were detected in the same rocks, Brocks *et al.* (1999) proposed that Eukarya may date as long ago as 2.7 billion years. The fact that putative traces of Eukarya and Cyanobacteria are found in the same rocks is interesting, as eukaryotic membranes contain cholesterol, which requires fairly high amounts of oxygen for its synthesis. One could then imagine that the local production of oxygen by Cyanobacteria was used by Eukarya to produce their membrane cholesterol. However, this appealing hypothesis is ruined by a recent article in Nature (Rasmussen *et al.*, 2008) (but see also Rashby *et al.* (2007)), which shows that the biomarkers found in these ancient rocks probably entered the rocks after their formation. The oldest evidence for Eukaryotes is thus found at 1.78-1.68 billion years ago, and for Cyanobacteria at 2.15 billion years ago.

Photosynthesis is 2.7 billion years old.

Biomarkers were also used to characterise an ecosystem that produced rocks from a 1.64 billion year old basin in Australia. Brocks *et al.* (2005) found molecules diagnostic of Chromatiaceae, a group of gamma-Proteobacteria, and of Chlorobiaceae, from the Bacteroidetes/Chlorobi group, which indicates that this particular ecosystem was mainly anoxic. This finding is consistent with the idea that oxygen remained fairly low until later than 1.64 billion years ago: some environments thus remained quite protected from oxygen.

2.4.3 Isotope ratios

Another kind of marker is found in isotope ratios. Atoms come in different isotopes, that depend upon the number of non-charged particles, the neutrons, that they contain. For instance, carbon atoms are found in three different flavours, ^{12}C, that contains 12 nucleons (6 neutrons and 6 protons) and makes for about 99 percent of all carbon, ^{13}C, and ^{14}C that decays in a few thousand years. Volcanic rocks can be dated by analysing the relative frequencies of different isotopes, notably in the couple uranium-lead. Other geologic strata can be dated relatively to these absolutely dated volcanic strata: this is how geologists can tell the age of a rock.

Biological reactions tend to prefer lighter atoms: living matter is therefore enriched in ^{12}C compared to ^{13}C and ^{14}C. When this living matter fossilises, it produces rocks enriched in ^{12}C: this makes it possible to establish the biogenicity of ancient rocks by measuring their $^{13}C/^{12}C$ ratio. Accordingly, Mojzsis *et al.* (1996) measured the $^{13}C/^{12}C$ ratio in Greenland rocks they estimated to be 3.8 billion years old, and found the depletion in ^{13}C characteristic of biological origins: they thus concluded that they had discovered the earliest traces of life. This result is debated however, because the rocks that were used for this measurement have been altered to the point that their dating is uncertain (Eiler, 2007). Another measurement by Rosing (1999) nonetheless also finds a $^{13}C/^{12}C$ ratio compatible with a biogenic origin in 3.7 billion year old rocks from Greenland.

Carbon isotopes have also been used to date the appearance of a particular metabolism, methanogenesis. Ueno *et al.* (2006) measured carbon isotope ratios in fluid inclusions in rocks from the Pilbara Craton in Australia, thought to have been deposited more than 3.46 billion years ago. Because the $^{13}C/^{12}C$ ratio in the embedded methane is consistent with a biogenic origin, and because abiotic processes would also have produced other gases that have not been found in these inclusions, the authors concluded that methanogen archaea must be at least 3.46 billion years old.

Methanogenesis is more than 3.46 billion years old.

Other isotopic ratios offer insight into the early Earth. For instance, ^{18}O and ^{30}Si have been used as palaeothermometers of the ocean. Notably, Robert

et Chaussidon (2006) estimated that during the last 3.5 billion years, average oceanic temperatures decreased from about 70°C to about 20°C today. Their reasoning was based on the fact that at high temperatures, these atoms are more easily disolved in sea water. Consequently rocks that formed in a sea water at high temperature tend to be depleted in ^{18}O and ^{30}Si compared to rocks that formed in colder sea water. When they analysed a sample of rocks spanning the last 3.5 billion years, and after discarding some rocks that may have been altered by hydothermal fluids, they found that their two markers were in good agreement and argued in favour of a decrease in oceanic temperatures. As temperature is a major parameter affecting living organisms, these results have important implications for the evolution of life.

Still another atom whose isotopes are useful for the study of the early Earth is found in sulphur. Sulphate-reducing bacteria combine hydrogen with sulphate to produce hydrogen sulphide. In doing so, they tend to show some preference for a particular sulphur isotope, the lighter ^{32}S compared to ^{34}S. Shen *et al.* (2001) studied 3.47 billion year old baryte rocks in Greenland, that are rich in sulphate. Their reasoning was that if sulphate reducing bacteria were present at the time, they probably reduced some of the sulphate that gave rise to the baryte rocks. Indeed they found some microscopic sulphide inclusion, and measured their $^{34}S/^{32}S$ isotopic ratios, to estimate that these were consistent with their production by sulphate-reducing bacteria. This means that this particular metabolism may be at least 3.47 billion years old. Moreover, sulphate is more abundant in aerobic milieux than in anaerobic ones, so that sulphate reduction must have been more important when oxygen started rising. Accordingly, Canfield *et al.* (2000) found an increase in sulphur fractionation between 2.75 and 2.2-2.3 billion years ago, consistent with the record of oxygen concentration. More recently, Bekker *et al.* (2004) measured sulphur isotope ratios in rocks from South Africa, and concluded that by 2.32 billion years ago, oxygen had reached at least 10^{-5} times its present level, whereas a few hundreds of million years earlier, oxygen was nearly absent from the atmosphere. Then a few hundred million years later, a new oxygenation episode took place. Between 0.58 and 0.55 million years ago, oxygen rose again, simultaneously with the first appearance of multicellular animals, during what is known as the Ediacara period (Fike *et al.*, 2006). This coincidence may be meaningful, as multicellular animals all need oxygen to live, and have developed a range of exquisite systems for providing oxygen to their cells, from gill slits and arterio-venal circulation to insects trachea. It is thus not unreasonable to assume that only when oxygen levels reached a certain threshold could animal multicellularity appear.

Oxygen has had a huge impact on life and on Earth.

As a matter of fact, it is believed that oxygen had other spectacular influences on animal life. Notably, Ward *et al.* (2006) proposed that colonisation of aerial, and not aquatic, ecosystems by arthropods and vertebrates occurred in

two phases, each triggered by increases in O_2 concentrations. A few million years later, about 0.3 billion years ago, in the carboniferous, oxygen reached nearly twice its present-day level, which is believed to be the reason why giant animals like a sea-scorpion 2.5 meters long (Braddy *et al.*, 2008), a dragonfly 0.75 meter wide (Dudley, 1998), or an amphibian 2 meters long (Dudley, 1998), are found in the fossil record of this period (Berner *et al.*, 2000): with higher partial pressures in oxygen, larger animals would have been able to supplement their cells with sufficient amounts of oxygen.

2.4.4 A sum-up of some insights from geological studies

Several types of indices suggest that life may be more than 3.5 billion years old (Fig. 2.2). Several important metabolisms have even been dated to be more than 3.4 billion years old: methanogenesis is dated at 3.46, and sulphate-reduction is dated at 3.47 billion years ago. Not surprisingly, these two metabolisms work in anaerobic environments. Oxygenic photosynthesis, the cyanobacterial metabolism that changed the face of the Earth by injecting en masse oxygen in the atmosphere, must be at least 2.4 billion years old.
Although Cyanobacteria clearly changed the Earth environment, it took quite some time for oxygen to reach its present level. Apparently, by 2.2 billion years ago, oxygen was more than 10^{-5} times its present level, but it reached its present level only around 0.6 billion years ago (Scott *et al.*, 2008). This is in agreement with the inference of the presence of two anaerobic groups, Chromatiaceae and Chlorobiaceae, 1.64 billion years ago. The date when oxygen finally reaches its present values coincides with the first appearance of multicellular animals; when it peaks to about twice present values coincides with the appearance of giant arthropods and giant amphibians. Oxygen has had a tremendous impact on the evolution of life, and many interesting studies remain to be done to study the link between this molecule and life history.

Robert et Chaussidon (2006) propose that oceanic temperatures were 70°C 3.5 billion years ago. This would suggest that early organisms lived at high temperatures, and progressively adapted to temperatures that are now met on the Earth. Article 6 shall detail a little bit what genomes have to say on this expectation.

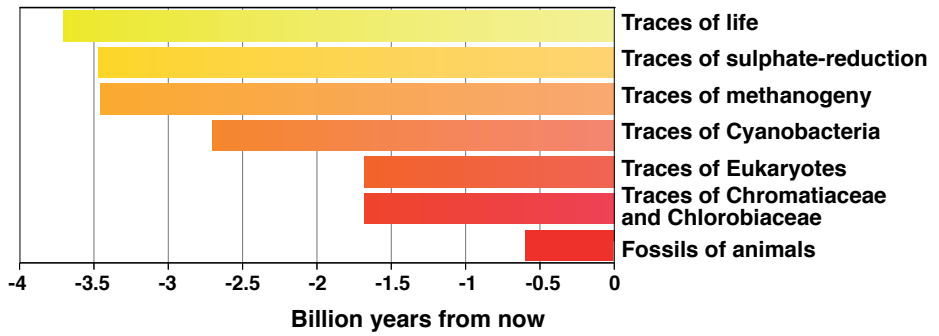


Figure 2.2: *Chronology of some events in the history of life.*

2.5 The historical content of extant organisms

Digging for information about the history of living beings in rocks seems a fairly natural quest, as rocks are full of fossils. As we have seen in the previous section, fossils of all sorts show that life has existed for a long time, but they are not very generous about the early phylogeny of living beings and how they came to be the way they are. Rocks more than 3 billion year old are very rare: an attentive reader may have noticed that studies that I reported in the former section all came from Australia, South Africa or Greenland: few other places harbour very ancient rocks (hopefully, climate change may grant us access to new areas of interesting rocks). Moreover, they have often been very altered as eons passed. All in all, they only give partial and shaky, but key, information.
Other documents of evolutionary history can be found in living organisms themselves: their morphology, their characteristics, but also their genome deliver information regarding evolution. Like rocks, extant species carry scars of ancient events in their bones and genes.

2.5.1 Morphological data and homology

Comparing extant species permits to infer their phylogeny.

The arrangement of similarities and differences between organisms betrays how they evolved from common ancestors. For instance, bonobos and chimps are, to our eyes, nearly identical creatures, although one is more affectionate than the other. This high similarity reflects their very recent divergence from a common ancestor: both species had little time to accumulate morphological differences one from the other. A slightly longer time in the past, chimpanzees and bonobos also share a common ancestor with humans, which is again shown by the large similitude between these African apes and their naked cousin. One can go on like

that and assemble a tree of life by gathering more and more distant cousins until no more living species is left that has not been invited to the family dinner. Such a "pilgrimage to the dawn of evolution" has been described by Richard Dawkins (2005), although for some reason he chose humans as focal species instead of bonobos.

Rather than using eyeball estimations of overall similarity as estimates of relatedness, evolutionary biologists rely on well defined methods to sort out species trees. One important step notably is to define *characters* and *states*, by individualising parts of the body. For instance, the forelimb in mammals may be such a character, as it can be easily individualised from the rest of the body. Then, for this character, states could be arm, wing, or flipper. Although the example I chose is rather straightforward, usually a major problem is to identify such characters. Why do we think it reasonable to consider that there were transitions between arm and flipper and wing? Said differently, why do we think that all three did not appear independently of each other but evolved from a single ancestral state in the ancestor of all animals under study? Said shortly, why do we consider that they are *homologous*?

Two characters are homologous if they evolved from a single ancestral character.

The definition of homology in evolutionary studies is of crucial importance. If one wants to infer the evolution of a character, one needs to recognise properly this character, under all the costumes it has disguised itself in primates, bats, seals, or any species one is interested in. For morphological characters, homology can be identified by looking at position, shape, time of occurrence in life history, or more recently pattern of gene expression. If there are lots of similarities between two characters in different species, it seems more reasonable to assume that these common points result from common descent, *i.e.* are homologous, than from parallel appearances.

High similarity suggests homology.

If morphology may provide valuable information to build a phylogeny of great apes, mammals or even vertebrates, it gets quite uncomfortable when one wants to compare a plant with a hummingbird, a bacterial parasite with its insect host, or a whale with an archaea living in the stomach of a cow: homologous characters are uneasy to find and their evolution is, to say the least, complex. Nonetheless, a proper tree of life should include all of life, if possible as a faithful depiction of how they evolved from a common ancestor. Non-morphological characters are thus much needed, and can be found in genomes.

Only genomes permit to reconstruct a tree of life.

2.5.2 Sequence data

Genomes are linear molecules made of four different types of sugars (nucleotides), named adenine, cytosine, guanine, and thymine, or, more shortly, A,C,G and T.

Reading the sequence of a genome therefore amounts to repeating A,C,G and Ts a large number of times in a specific order. In practice phylogeneticists do not compare genomes in their totality, although they could, but for practical reasons select a few *semantids* (Zuckerkandl et Pauling, 1965b), which most often are simply genes, *i.e.*, approximately, functional segments of the genome sequence that can be transcribed and possibly translated. In fact, to build a phylogeny, all one has to find is portions of genome sequences that can be compared between species, *i.e.*, portions of genomes that are homologous.

When comparing semantid or genome sequences to build a phylogeny, characters are single nucleotides. Homology between these sequence characters is based on concepts inherited from morphological analyses. It is determined based on overall sequence similarity: if two sequences show a lot of similarity, it is unlikely that they arose independently of each other, and therefore they must be homologous. Determining overall sequence similarity is achieved through sequence alignment, a procedure whose aim is to align sequences with respect to each other so that their overall sequence similarity is maximised. In the end, if the maximal score achieved is such that homology is most probable, sequences can be used for phylogeny reconstruction. Several heuristics exist that can align sequences together (Thompson *et al.*, 1994; Notredame *et al.*, 2000; Edgar, 2004; Löytynoja et Goldman, 2008), or whose aim is to quickly search a database for sequences similar (and then possibly homologous) to a query sequence (Altschul *et al.*, 1997).

Other sequence-based characters can be used, like the presence/absence of a gene in the genome of an organism. In such a case, a character has only two states, "present" or "absent". However this type of analysis is usually less precise than analyses that use sequences in their entirety. Actually, sequence data offer several advantages compared to other data for evolutionary reconstructions.

First, it is very easy to obtain sequence data. As seen in television series, a single hair can be used to extract DNA and sequence a gene from it. Then, it is easier to establish homology for sequence data than for morphological data, up to the point that a computer can do it. If there is a gene of interest in the Bonobo genome, it is not that hard to find a homologous gene in Human, for instance. If my bonobo gene is 500 characters long, I can decide that whatever human gene has a sequence more than 70% identical (350 characters in common) is a homologous gene. A quick computation shows this 70% identity criterion is conservative: in DNA, there are 4 possible states at each site. 70% identity means that among the 500 characters, more than 350 are identical between Human and Bonobo. The probability that at least 350 sites are identical in 500 sequences

can be computed as follows:

$$P(\textit{more than } 350 \textit{ identical sites in } 500 \textit{ sites}) = \sum_{i=350}^{500} C_{500}^{i} \times (\frac{1}{4})^{i} \times (\frac{3}{4})^{500-i}$$
$$= \sum_{i=350}^{500} C_{500}^{i} \times \frac{3^{500-i}}{4^{500}} \approx 7 \times 10^{-99}$$

The probability that the Bonobo and Human sequences are identical at 70% by chance alone (in other terms, convergence) is so small ($\frac{1}{7\times10^{99}}$) that it is nearly certain that two sequences this similar descend from a single ancestor. Then I can run a program that scans the whole human genome and reports all sequences that are more than 70% identical to my gene of interest. All these sequences, according to my conservative criterion, are homologous genes.

It is much more difficult to find objective criteria to establish homology for morphological data: as said previously, to establish the homology of an organ, one can consider its three-dimensional structure, its embryological origin, the genes that are expressed during its formation, the developmental phase during which it is formed... Simply gathering all this information requires a lot of tedious work. In the end, if lucky, a researcher may have assembled a large body of data, some of which may be advocating homology, others not. From these, the researcher has to decide whether he can confidently assert homology or if he is unsure, and no simple automatic method can help him.

Consequently, using sequence data is much more practical than using morphological data; another advantage of sequence over morphological data will be presented in section 2.6.3.

Once characters have been gathered, either morphological or sequential, a crude way to reconstruct the phylogeny of species would be to use criteria akin to "birds of a feather flock together": animals sharing the highest number of states are expected to be the most closely related. More elaborate tools have been developed, and can be grouped under the flag of inferential statistics.

Comparing genomes requires statistics.

2.6 Statistics for inference

People who are fond of genealogy can search parish records for information about their ancestors. This way they can discover the names and jobs of their grand-grand-grand-grand-father, for instance. Another way to know more about their forebears would be to find their brothers, sisters, and cousins, ask them to spit in a tube, and then sequence their genomes. Then, by carefully analysing the

genomes of his kin, and using his knowledge of how genes are passed from one generation to the next, a genealogist might be able to reconstruct characteristics of his glorious antecedents, and know if they could curl their tongue into a tube and if they had wet earwax, for instance.

The phylogeneticist is faced with roughly the same choices: either he can search the fossil record to find information about his ancestors, and do paleontology, or he can look at his genome and the genomes of his cousins (of their morphology), and do comparative genomics (or comparative anatomy). The only difference here is the timescale. When the genealogist is interested in decades or at best centuries, the phylogeneticist is interested in a history that spans thousands to billions of years. The phylogeneticist cousins consequently may be more hairy, much bigger, much smaller, in every respect very different from him. Despite possibly huge differences, he can study their morphology or, if morphologies are different to the point that enough homologous characters cannot be recognised anymore, their genomes. Hopefully, he may come up with a satisfying depiction of the history of their family, and infer characteristics of their ancestors. To do so, he needs models of morphology and sequence evolution, and a bit of statistics.

In this section I am going to explain what I mean by "model" and statistics, introducing these concepts with a simplistic example. Then I will rapidly present models of evolution, and in subsequent sections I will present what such models of evolution can tell us about the history of life on Earth, and combine this knowledge with what we've learnt from geology.

2.6.1 A leak example

Building hypotheses to understand an unexpected pattern

Let's assume I have a small problem of water overconsumption. When I look at the water that is used on a daily basis in my apartment, I have the impression that some days too much water is wasted: I suspect that someone, let's call her M, may sometimes forget to properly close some tap before going to bed. However, I cannot charge her unless I am really sure, and I also want to know how frequently she forgets to turn off a tap, to know how big a punishment she deserves. All I can do is look at the daily water consumption for a set of dates, and then try to prove, based on these data only, that M did commit a water crime. To do this, I can compare two cases: either taps are always properly closed before going to bed and the days where we spend a lot of water are the fruit of an expected variation in normal water consumption, or there are indeed some days where a tap is not correctly turned off, which results in a waste of water. These two cases correspond to two types of *processes*, one where the water consumption derives

Two hypotheses: either some taps are sometimes left open, or not.

from only random variation around some average value (A), and one where it derives from the random variation around some average value, plus sometimes an additional quantity of wasted water (B). To catch M, I need to prove that (B) happened, *i.e.* generated the data, not (A). To do so, I can compare what daily water consumptions process (A) would produce and compare it to the real data, and do the same for process (B). If process (B) produces data that more closely resembles the real ones than process (A), then I can joyfully charge M.

Now comes the problem of knowing what data processes (A) and (B) would engender. First, I could make "physical" simulations. For instance, I could build a copy of my apartment, use water in the copy apartment as similarly as possible as we do in the real apartment, and then in addition sometimes let a tap open for a night. I could try to let the tap open 0 (model (A)), 1, 2, 3... n (model (B)) times, and see under which number of times water is spent in a similar way as observed in the real data. However, this simulation technique would be tedious and difficult to implement.

A model is a simplified interpretation of the reality.

Instead, I can use computer simulations based on a *probabilistic model*. This probabilistic model needs to incorporate the normal variation in daily consumption, but also the occasional forgotten open tap. A simple model for daily variation around an average quantity is provided by the normal, or Gaussian, distribution. This distribution is defined by two parameters, the mean μ and the standard deviation σ. For the occasional forgotten open tap, I assume that with a probability p, a tap has not been turned off before the night, so that some quantity Q of water is wasted. Obviously, this Q parameter is a poor approximation of the reality: one does not expect that when a tap is left open, exactly Q litres are wasted, whatever the duration of the night and the flow. However, when building models of a real process, simplifications need to be made, and I hope in these circumstances this simplification will not do much harm to the applicability of my model. This model therefore totalizes 4 parameters. With such a probabilistic model, I can run simulations *in silico*, which is much easier and faster than in the real world.

Confronting the data

If a model can generate realistic data, it is a good model.

Now that I have my probabilistic model, I want to know how many times M forgot to turn off a tap (with the possibility that this number of times is 0, process (A)). To this end, I simulate data under my model with different values for the parameters μ, σ, p and Q. Then I compute distances between real data and simulated distributions. For real data, I have monitored water consumption for 365 days. To compare real and simulated distributions, I use the following protocol:

1. First I order the 365 real values r_i on one hand , and the 365 simulated values s_i on the other hand

2. Then I compute $Distance = \sum_{i \in [1..365]} |r_i - s_i|$

The distance that I compute is the simplest I could think of. It may not be the best distance possible: for instance, this distance might be biased in some way, or be little discriminating. Further studies should be made to ensure that my *statistical estimator* is not too bad.

I simulate distributions with the following parameters:

- μ: all integers in [20;80], by increases of 2
- σ: all integers in [0;40], by increases of 4
- p: all real numbers in [0;1] by increments of 0.1
- Q: all integers in [5;80] by increases of 5

For each set of values, I run ten simulations.

In the end I choose as my best estimates for μ, σ, p and Q the values that produced the distribution closest to the true one.

When I follow the above protocol, I find that the best estimates of my parameters are as follow:

- μ: 60
- σ: 4
- p: 0.1
- Q: 20

This means that our best model according to our estimator predicts that on average, we consume 60 litres per day, with a standard deviation of 4 litres, and that M has forgotten about 36 ($p \times 365$) times to switch off the tap, which may have cost the loss of as much as 730 ($Q \times p \times 365$) litres in total. The fit of our model is not too bad, as shown when I superpose the densities of the two distributions Fig. 2.3. We can notably observe that the real distribution shows some bumps on the right side, as the simulated distribution. These bumps most likely come from leaks.

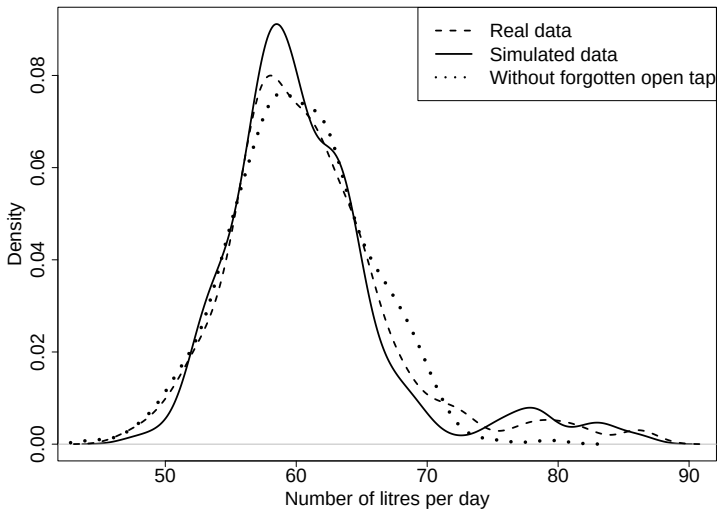


Figure 2.3: *Simulated distribution with best fit to the real distribution, and distribution obtained under the model without occasional open tap.*

It is probable that if I had tried more values for the parameters of my model, I would have obtained a better fit. Moreover, I do not know whether there is a huge difference between choosing these specific values for the parameters and choosing other close values, or even choosing a model where $p = 0$ (process (A)): perhaps other values provide a fit nearly as good as these values, in which case the weight of evidence for these would be very weak. To get a qualitative idea as to whether process (A) may have produced the real data, I have represented the distribution closest to the true data when $p = 0$ in Fig. 2.3. Although it is not that far from the real distribution, it does seem to be not as good as the best distribution when $p \neq 0$: notably, it does not produce bumps as in the real distribution. It would be probably safe to make sure that the model with $p = 0$ really cannot build a distribution as good as when $p \neq 0$, possibly with more simulations or by increasing the sample size, but I am going to trust this inference: after all, not much is at stake.

When confronted with these results, M confessed that she had forgotten to close a tap 24 times during the last 365 days. This shows the advantage of statistics for running a household.

2.6.2 Inferential statistics

The leak example showed that, when confronted with some unexpected data, one can make hypotheses about the process that generated them, build a model based on these hypotheses, and test how well this model (the hypotheses) fits the data, according to some estimator. This whole procedure can be referred to as inferential statistics.

As in our example, living species and their characteristics are the result of a process that occurred through time, evolution. As in our example, we have no way of knowing the process with certitude, because we have no time machine. As in our example, we can resort to inferential statistics to make hypotheses and confront them to data.

Inferential statistics need 4 elements:

1. The data. In the leak example, data were composed of quantities of water spent per day; in biology, these data could be the presence/absence of some characters (a cell nucleus for instance), an ecological factor (the optimal growth temperature for instance), sequences, *etc*...

2. Hypotheses on the process that generated the data. When faced with data, a statistician elaborates some hypotheses of how they came to be the way they are, hypotheses about the process that generated the data. For the leak example, one could think of two possible processes (A) and (B).

3. A model. Once the hypotheses have been enunciated, they need to be translated into a mathematical model. Such a model needs to incorporate the most important aspects of the real process: an optimal balance between realism and tractability needs to be found.

4. An estimator. Once the statistician has a model of the processes he thinks may have generated the data, he needs to find a way to analyse how closely his model fits the data. In the leak example, I resorted to simulations and computed distances between these simulations and the true data. Other estimators have been found and studied in a wide range of problems, so that when a particular estimator is used, one knows its characteristics.

Modelling evolution to reconstruct history.

When applied to biological data, the aim of inferential statistics is to reconstruct past history, either past speciations to estimate the phylogeny, or past events that occurred between speciations, to estimate the process of evolution. We have already presented the data that phylogeneticists could use to infer the history of life from extant organisms. With these data in hand, phylogeneticists can use the lenses of models of evolution to look into the past.

2.6.3 Models of evolution

Two different types of data impose two different types of models. Some characters show a limited number of states: these are discrete characters: for n states, there are n^2 possible transitions, including transitions from state i to itself. Other characters do not have a limited number of states (for example, the size of an organ or the number of hair on the back of a drosophila), but their evolution can still be modelled probabilistically. I will not detail this second kind of models as I have not used them; further details about these can be found notably in Felsenstein (1973); Pagel (1999); Pagel *et al.* (2004). In the next few lines, I briefly present models that have been used for discrete characters with a fixed number of states.

Similar models of evolution can be used for morphological and sequence data. For sequence data, the number of states is clearly defined: for DNA, there are 4 states. For morphological data, this number will vary: if forelimb is the character of interest and its observed states are arm, wing, and flipper, this number is 3. In both cases, models of evolution can be devised. Early evolutionary biologists had (and for a great part we still have) no idea how morphological changes occurred, and unfortunately did not have access to sequence data. Consequently, very few hypotheses could be made about the processes that generated morphological diversity. The corresponding models were thus necessarily very simple, and very subjective, giving *a priori* more probability to a particular transition than to another. Moreover, early models were estimated by hand, as computers were not available yet. This (although this is clearly not the only reason) imposed a further constraint on model realism: when a character had different states in two species, by default, one would count only one transition event, even if there may have been several transitions in chain, especially if the two compared species have diverged a long time ago, or if the character under study is little constrained and undergoes many transitions. In consequence, such a model, named *parsimony*, is known to be subject to bias when lots of events have occurred, because it does always underestimate the true number of transitions, as has been repeatedly shown on sequence data (Felsenstein, 1978; Hasegawa et Fujiwara, 1993; Kuhner et Felsenstein, 1994; Tateno *et al.*, 1994; Huelsenbeck, 1997; Guindon et Gascuel, 2003). Instead, especially on sequence data, people now use more flexible models, that can estimate that a character has undergone several transitions even if these transitions have left no observable trace, and that rely on an explicit probabilistic model, where transitions are associated with probabilities (Jukes T.H., 1969; Dayhoff *et al.*, 1972; Kimura, 1980; Felsenstein, 1981; Lanave *et al.*, 1984; Hasegawa *et al.*, 1985; Tamura, 1992; Jones *et al.*, 1994; Whelan et Goldman, 2001; Le et Gascuel, 2008). These transitions are parameters of the model, which can be statistically inferred. Models of sequence evolution have been thoroughly detailed in Felsenstein (2004); Yang (2006), for instance, or in the theses of Galtier (1997); Guindon (2003). More details will also be provided

in section 2.7.2.

A model of evolution contains transition probabilities and a phylogeny.

The model of character evolution should not only account for the process, with the transition probabilities, it should also account for the pattern, the path that has been taken by evolution to yield the observed state distribution among species. This pattern is usually modelled by a bifurcating acyclic rooted graph, and represents a species phylogeny (Fig. 2.4). This graph is composed of nodes and branches (according to the phylogenetics vocabulary), where the root node represents the ancestor of all organisms found in the tree, internal branches join two ancestors together, and external branches join a sampled species with its most direct ancestor. In practice, one can impose a known species phylogeny, or estimate the species phylogeny jointly with the other parameters of the model of evolution. However, in such cases, one character is not enough to estimate it all. When both species tree and character evolution are to be estimated, a large number of characters are needed.

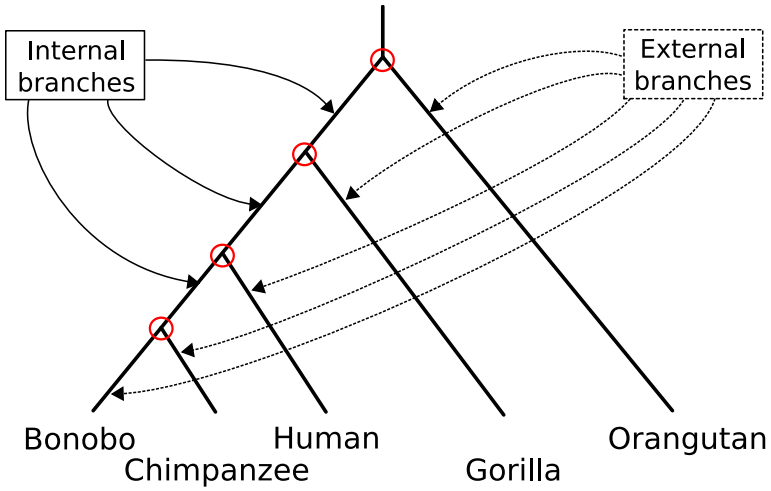


Figure 2.4: *Example of a species tree. The species of interest are bonobo, chimpanzee, human, gorilla and orangutan. Internal nodes are circled in red. The root of the tree is the uppest node in the tree.*

Sequence data are easier to model and contain more information than morphology.

Here the second advantage of sequence data over morphology is very obvious, as the number of sequence characters can be very large, and as each character is fairly similar to the next one: a limited number of parameters may fit a large number of sequence characters. On the contrary, morphological data are difficult to acquire in large quantity, and may be less easy to model with a limited set of parameters, as from one character (for example: the forelimb) to the next (for example: the presence of mammary glands), states may differ (in our examples

of characters, one may have a dozen of states and the other only two). As it is more difficult to properly estimate parameters of a model when data are scarce, using sequence data to reconstruct evolutionary history is a much more reasonable endeavour.

Because it is easier to study sequence than morphology evolution, models of sequence evolution have recently reached a degree of sophistication that models of morphological evolution never reached. It is now known which kind of substitution among A, C, G, T is usually more frequent than the other one, that the model of substitution can change depending on its position in the alignment, or depending on the time in its history. More formal presentations of models of evolution can be found in articles 4 and 8. Heterogeneity in models of evolution through time has been tackled in section 2.7.2 and articles 4 and 7, where different models of evolution can be used on different branches of a phylogeny. Heterogeneity in models of evolution between sites has been tackled in 8, where different phylogenies are associated to different portions of a single semantid.

2.6.4 Estimators

Once a model of character evolution has been defined, to infer both species tree and character evolution, all that remains to be found is an estimator. One could use simulations and compute distances to the real data as was done in the leak example. Instead, because in phylogenetics simulations would have no advantage over them, more classical estimators are used. The first estimator that was used was in the context of parsimony models, and is named maximum parsimony: this estimator posits that the best model (here I understand model as the phylogenetic tree only), is the model that supposes the smallest total number of transitions between states. It was first applied to molecular data by Edwards et Cavalli-Sforza (1964), and an improved algorithm was devised by Fitch (1971) a few years later. It may work finely when the characters under study have undergone few transitions; however, some characters evolve very fast, so that their true history will differ from the most parsimonious one. Therefore maximum parsimony should be used with caution, and more flexible estimators be preferred. Mainly three such estimators that rely on explicitly probabilistic models of transition between sites have been used in practice, although mostly on sequence data. These estimators are *minimum evolution*, *maximum likelihood*, and *Bayesian integration*.

Minimum evolution (first exposed by Cavalli-Sforza et Edwards (1967)) is related in philosophy to maximum parsimony, as it likens the true evolutionary history with the one that supposes the smallest number of transitions. In practice, very fast heuristics have been found that produce fairly good results ((Saitou et Nei, 1987; Studier et Keppler, 1988; Gascuel, 1997), see for instance (Guindon

et Gascuel, 2003) for their efficiency). However, this type of method does not permit to accurately estimate parameter values other than the tree topology. For instance, transition probabilities cannot be inferred with this estimator. This is a drawback as knowledge of the process, or an estimate of the process, can improve phylogenetic reconstruction.

Maximum likelihood and Bayesian integration are the best estimators in phylogeny.

Instead of a quantity of evolution, the focus of maximum likelihood and Bayesian integration is probability. In practice, they are much slower than methods based on minimum evolution, but are more precise, and permit to use better models of sequence evolution, as other parameters than the tree topology can be estimated. Maximum likelihood takes as an estimator of the true evolutionary history the one that permits to maximise the probability of the data (the likelihood of a model is proportional to the probability of the data given a model). This is in essence very similar to the poor technique that I have used in the leak example: if an infinite number of simulations were done, the maximum likelihood model would be the one that produces the real data most often among all models considered. The algorithm used to compute the likelihood of a phylogenetic tree for discrete characters proposed in 1981 by Joseph Felsenstein (Felsenstein, 1981) does not require simulations however but uses an analytical formula, and is applicable to any type of data that can be described with a limited number of states. Formulae for computing the likelihood of a phylogenetic tree are given in articles 4 and 8.

Bayesian integration is different from all estimators discussed so far, as it does not provide a point estimate of the best model (although it is possible to extract point estimates from the result of an analysis by Bayesian integration). It takes instead a more cautious approach, by acknowledging that any estimation is associated with a certain amount of uncertainty: it produces a probability distribution over the models of interest, which can then be summed up. The probability used by Bayesian integration is posterior probability, not likelihood: while likelihood is the probability of the data given the model (or is proportional to it), the posterior probability is the probability of the model given the data. Differently put, the model with the highest posterior probability is the model that most probably generated the data. This allows one to naturally compare models with each other: if model A has a posterior probability of 0.09 and model B a posterior probability of 0.03, this means that model A is three times more probable than model B. On the contrary, if model A has a likelihood of 0.09 and model B a likelihood of 0.03, one *cannot* say that model A is three times more probable than model B, but could say that data are three times more probable under model A than under model B. In this case, one can say that model A is three times more likely than model B, a distinction in terms due to Fisher (1922). Let's look at some formulae to better understand the difference between likelihood and posterior probability.

The likelihood $L(M|D)$ of a model M is proportional to the probability p of the data D given the model M:

$$L(M|D) = k \times p(D|M)$$

Here, we consider that the proportionality constant k is 1:

$$L(M|D) = p(D|M)$$

The posterior probability $PP(M)$ of a model is the probability of the model given the data:

$$PP(M) = p(M|D)$$

In molecular phylogenetics, M corresponds notably to the set of transition probabilities between states, and the phylogenetic tree. D corresponds to the sequences under study.

When one considers two models A and B, one can compute their likelihoods $L(A) = p(D|A)$ and $L(B) = p(D|B)$. These two likelihoods are probabilities, but they are not from the same probability space. On the contrary the posterior probabilities of models A and B are $PP(A) = p(A|D)$ and $PP(B) = p(B|D)$, from the same probability space. As a consequence, for a given dataset, the sum of posterior probabilities for all models is 1: $\sum_M PP(M) = \sum_M p(M|D) = 1$, but the sum of likelihoods for all models is undefined.

More precisely, posterior probabilities permit building a *probability distribution* over all models, whereas likelihoods do not. Using likelihood, one could also get a probability distribution, but a probability distribution over all possible data, for one model: not interesting for a statistician, who usually only has got one dataset, and wants to find the best models.

A probability distribution is a well-defined mathematical object, so powerful techniques exist to work with them. It is notably possible to sample from them "smartly", *i.e.*, here, to avoid wasting time sampling models that have a very weak posterior probability, and not miss models that have a very high posterior probability. Indeed, posterior probability distributions often cannot be fully explored, because there are too many possible values. This is notably true in phylogenetics, where the number of possible trees is huge. Instead, one samples models, using techniques such as Markov chain Monte Carlo (MCMC) (Metropolis *et al.*, 1953). If run infinitely, these techniques guarantee that the set of models sampled will be an unbiased sample from the full distribution. If run for a sufficiently long time, one can expect that the obtained sample will be very good. MCMC techniques work only for probability distributions: therefore to obtain model probability distributions, MCMC techniques can only be used with posterior probability distributions. Some authors have shown how to sample from the

likelihood function through MCMC, but as far as I understood they used MCMC on posterior probability distributions and then used a mathematical trick known as importance sampling (or importance reweighting) to transform the posterior probability sample into a sample of the likelihood function (Geyer, 1991; Kuhner *et al.*, 1995).

The application of Bayesian and MCMC methods to phylogenetics dates from the mid-nineteen nineties, with the pioneering articles of Rannala et Yang (1996); Yang et Rannala (1997); Mau et Newton (1997); Li *et al.* (2000). They have considerably gained in popularity since then, and will probably get more used as models of evolution become more elaborate. In this thesis, my work has used maximum likelihood techniques, although all models and algorithms that I used and developed can also be implemented in a Bayesian framework. In fact, likelihood and posterior probability are intimately related by Bayes' formula:

$$PP(M) = p(M|D) = \frac{p(D|M) \times p(M)}{p(D)} = \frac{L(M) \times p(M)}{p(D)}$$

In this last formula, one can see that the posterior probability of a model is proportional to the product of the model likelihood and of a prior probability $p(M)$ associated with the model, arbitrarily defined by the user of a Bayesian program. Another term is found in $p(D)$, the probability of the data. This last term is difficult to compute, and is usually not computed, so the posterior probability of a model can only be known up to a multiplicative term ($\frac{1}{p(D)}$). In practice, when MCMC techniques are used, this multiplicative term is cancelled out of the equations.

In some cases, the most likely models will also be models of highest posterior probabilities. This is notably true when the prior probabilities $p(M)$ do not differ between models, *i.e.* $\forall M, p(M) = c$, *with* $c \in [0;1]$ *constant*. In such cases, $PP(M) = p(D|M) \times c = L(M) \times c$. However, if the user of a Bayesian program has some knowledge on which models are more probable than others, he can affect different prior probabilities to the models. This may result in differences between the most likely models and the models of highest posterior probabilities.

These differences are where proponents of maximum likelihood methods (frequentists) and proponents of Bayesian integration (Bayesians) disagree, because prior probabilities influence the result of an analysis. Frequentists think that it is bad to influence what the data have to say, and Bayesians answer that their method has more power, because it incorporates extra knowledge than what is only found in the data. There are however cases where scientists have no idea what prior distribution should be used. In such cases, Bayesians would suggest using a uniform distribution; however frequentists answer that using a uniform

distribution is far from an agnostic approach, as it amounts to supposing that all hypotheses are equally probable (Edwards, 1972). All would agree that much care must be devoted to assessing the impact of prior probability distributions on the result.

To me, Bayesian integration methods are attractive notably because they allow the statistician to handle models with more parameters than maximum likelihood. Indeed, the quantity of data one can study is limited: therefore an uncertainty is necessarily associated to the estimation of each parameter of a model. When there are lots of parameters, there is some chance that a few parameters will have a non-negligible amount of uncertainty. Relying on point estimates for these propagates this uncertainty to all other parameters. Instead Bayesian integration methods provide probability distributions for all parameters of the model: the distribution will be flat if there is a lot of uncertainty, or very pointy if there is a lot of signal in the data in favour of a particular value of the parameter. If a parameter value is uncertain, a maximum likelihood analysis will be highly uncertain; instead a Bayesian integration analysis will "integrate out" this uncertainty to estimate other parameters, and will therefore be more robust.

The alert reader will have noticed that all aspects of Bayesian integration that I praise are related to integration of uncertainty rather than to the use of prior probabilities (although these prior probabilities can be very useful in some models , e.g. (Rannala et Yang, 1996; Ané *et al.*, 2007)). As noted earlier, through importance sampling, it is possible to sample from the likelihood function instead of the posterior probability distribution. Such approaches lend robustness to likelihood-based approaches.

Sequence data, models of evolution and estimators have been used in conjunction to answer questions that geology and comparative anatomy could not address.

2.7 A short history of life on Earth, as told by genomes

Molecular phylogeny and geology are complementary.

In this section, I present some insights into the early evolution of life that were gained through the analysis of gene and genome sequences through models of evolution. Such analyses are complementary to geological studies. Geology provides punctual information about a particular environment at a particular time. The study of genomes in a historical framework gives another window into the past: for each genome analysed, new information about an ancestor is gained. As the number of genomes studied increases, the number of ancestors for which infor-

mation is available increases, and can help fill the gaps of the geological record. Moreover, the geological record can at best provide morphological data (and not really at the timescales that I have considered in my work), which, as I said in section 2.6.3 are less easy to analyse than sequence data. Overall, sequence data provide an unmatched way to study ancient evolution.

2.7.1 The three kingdoms

A cellular genome was first sequenced in 1995 (Fleischmann *et al.*, 1995); before that date, molecular phylogeny was often based on single gene sequences. One particularly important phylogeny based on sequences was even obtained without knowing the precise sequence of the studied gene, small subunit ribosomal RNA (rRNA) (Woese et Fox, 1977). Instead, the gene product, extracted for the organism under study, was submitted to an enzymatic treatment (a digestion), that cuts the molecule in little pieces. The way the molecule is cut depends on its sequence, therefore one can compute distances between the digestion patterns obtained for the rRNAs of different species, and extrapolate that this distance between rRNAs is a good estimate of distances between species. When Woese et Fox (1977) undertook their analysis, it was widely accepted that the primary division was between Bacteria, on one hand, and Eukarya, on the other hand. However, distances between rRNA sequences suggested that methanogenic bacteria were very different from the other ones, about as different from them as Eukarya. Woese et Fox (1977) concluded that methanogenic bacteria were not bacteria, and created a group for them, the Archaea. There were thus three different groups of species: Bacteria, Archaea, and Eukarya (Fig. 2.5). These three kingdoms have then been confirmed by the analysis of gene sequences, as efficient techniques revealing sequences in their totality have been developed (Sanger *et al.*, 1977). More representatives of the Archaea have also been discovered, which further convinced sceptics that there were three kingdoms (Forterre, 2007).

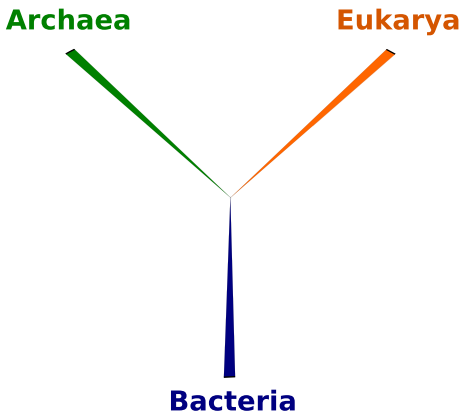


Figure 2.5: *Schema of a very simplified tree of life, with the three kingdoms Archaea, Bacteria and Eukarya.*

2.7.2 The root of the tree of life

The trichotomy between Archaea, Bacteria and Eukarya does not provide a way to pinpoint the root of the tree of life, the organism from which all extant organisms are descended, LUCA. Gogarten *et al.* (1989) and Iwabe *et al.* (1989) found an elegant way to root the tree of life. To understand it, one first needs to remember that LUCA is not the first living organism on Earth: LUCA is the Last Universal Common Ancestor. Many organisms may have lived before it, some of which may have left no descendant among extant organisms, others that would have been ancestors of LUCA. In the ancestors of LUCA, mutations occured; notably, some genes duplicated. With this in mind, they searched for gene families that were older than LUCA, and that had duplicated before its appearance. Each duplicate could then be used as an outgroup for the other one, and the tree of life be rooted (Fig. 2.6).

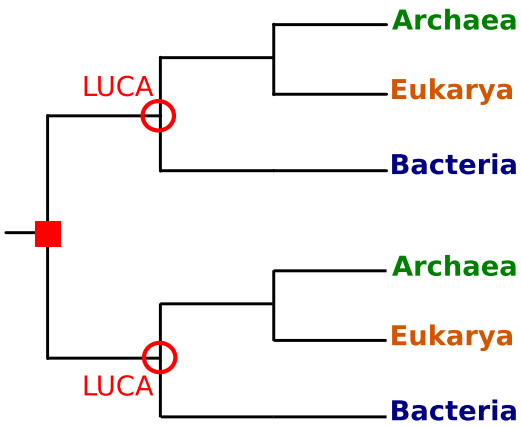


Figure 2.6: *The use of anciently duplicated genes for rooting the tree of life. The duplication event predating LUCA (red circle) is shown with a red dot. Such ancient duplicates suggest that the first speciation was between Bacteria and a group consisting of Archaea and Eukarya.*

It is unsure where the root of the tree of life is.

Analyses of anciently duplicated genes most often place the root between Bacteria and Archaea-Eukarya (Zhaxybayeva *et al.*, 2005). However, phylogenies obtained using these ancient duplicates have been questioned as too many substitutions may have affected their sequences (Forterre *et al.*, 1992; Philippe et Forterre, 1999). In this thesis, when inferring events close to the root, I therefore considered three possible roots, between each of the three kingdoms (articles 4, 6).

Few other ways have been proposed to root the tree of life. One possibility is to date some nodes of the tree and make the hypothesis that the substitution process is clockwise (Zuckerkandl et Pauling, 1965a; Huelsenbeck *et al.*, 2002; Kumar, 2005): under this hypothesis, the root of the tree should be the point equidistant from all extant organisms. However, the substitution process is rarely perfectly clockwise, especially when large evolutionary distances are considered, and better models of evolution that relax the clock hypothesis can be used to root a tree (Gillespie, 1984; Drummond *et al.*, 2006; Rannala et Yang, 2007). Although this method is interesting and deserves to be developed I shall not comment further on it.

In principle, the pattern of substitution might also be used to root a tree: under some models of sequence evolution, the direction of change has some importance (Yang et Roberts, 1995; Galtier et Gouy, 1998; Huelsenbeck *et al.*, 2002; Yap et Speed, 2005; Boussau et Gouy, 2006). However, the use of such models has been very rare, as it seemed less practical to search for a phylogeny using them (for more information see article 4).

All models of evolution assume that the substitution process follows a continuous-time Markov chain, *i.e.* that the next transition only depends on the present state, and not on the former state. This chain can be represented with the following matrix Q, showing instantaneous rates of transition (here also called substitution) q_{ij} between states i and j, in the order A, C, G, T:

$$\mathbf{Q} = \begin{pmatrix} - & q_{CA} & q_{GA} & q_{TA} \\ q_{AC} & - & q_{GC} & q_{TC} \\ q_{AG} & q_{CG} & - & q_{TG} \\ q_{AT} & q_{CT} & q_{GT} & - \end{pmatrix}$$

This matrix reads as follows: the instantaneous substitution rate from A to C is q_{AC}, from A to G q_{AG}, *etc.* $-$ is specified by the requirement that the columns sum to 0. From these instantaneous rates of substitution, one can derive probabilities of occurence for all possible substitutions during a time t by taking the exponential of the matrix Q:

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

If the matrix is irreducible, *i.e.* all states can be obtained after a given time whatever the starting state, then the chain has a stationary distribution, which means that if the chain is run for an infinitely long time on a long sequence, starting from any initial composition F_0, the end sequence composition will correspond to the stationary distribution Π:

$$\lim_{t\to\infty} P(t) \times F_0 = \Pi = \begin{pmatrix} \pi_A \\ \pi_C \\ \pi_G \\ \pi_T \end{pmatrix}$$

In addition, a model of substitution is said to be *reversible* if it satisfies the following equation:

$$q_{ij} \times \pi_i = q_{ji} \times \pi_j \qquad (2.1)$$

This means that if sequences are evolved at equilibrium using a reversible model, no matter how data are analysed, there is no way to tell in what direction evolution has occured: there is no way to find the root of a tree with a substitution model if sequences have been evolved under a reversible model. Conversely, had sequences been evolved with a non-reversible model of evolution, if the resulting data are analysed with a reversible model of evolution making the hypothesis that the process of evolution was at equilibrium, the signal for irreversibility is

ignored, and the root cannot be recovered.

In practice, most models of evolution that are used to reconstruct the history of sequences are reversible, not because it is believed that the real biological process of evolution is indeed reversible, but because it makes computations easier. Consequently it is usually hypothesized that the substitution process is at equilibrium, *i.e.* that sequence composition is the same all over the tree, and any signal for non-reversibility is ignored.

There are therefore two drawbacks to using reversible models of evolution: first the root cannot be identified based on sequences only even if there was a signal in the data for it, and second it is hypothesized that sequence composition has been constant throughout evolution, which is known to be wrong (for more discussion on it, see articles 3, 4, 6, 7).

Instead, one can use non-reversible models of evolution, that do not verify equation 2.1. In principle at least, these models permit one to find the root of a tree. In my work, I used models that permit inferring the root of a tree, but all substitution matrices that I used were reversible.

Indeed, a process of evolution can become non-reversible if it is made of several reversible processes combined. Notably, I have used particular models of evolution, called non-homogeneous or branch-heterogeneous models in this thesis, where different reversible substitution matrices are associated to different branches of the tree. Although this model is composed of reversible matrices, it is as a whole non-reversible, which may serve to pinpoint the root of a phylogenetic tree. I used this property to test several potential roots (article 4), but found that on the data analysed, the method was not powerful enough to identify a root without a doubt. This lack of power echoes recent works (Huelsenbeck *et al.*, 2002; Yap et Speed, 2005) that had used non-reversible matrices and had reached similar conclusions, but appears to be disappointing with respect to early expectations (Yang et Roberts, 1995).

These types of models, where different matrices are associated to different parts of the tree also have the advantage that they do not hypothesize that sequence composition has been constant throughout evolution; incidentally this characteristic was the main motivation behind their development. I therefore used such models to reconstruct sequence evolution from LUCA to extant organisms (articles 3, 4, 6).

Although these efforts based on models of sequence evolution have not been able to provide strong evidence in favour or against particular roots of the tree of life, I am not pessimistic on the possibility to find new ways to place LUCA. For

instance, one could combine non-homogeneous models of evolution with relaxed-clock models to benefit from both signals, but could also incorporate other types of information that permit to date a clade, like gene transfers (see article 10), and could analyse much more data than has been done up to now.

2.7.3 Primary endosymbioses

Most Eukarya harbour organelles (section 2.3), notably mitochondria, and chloroplasts. These organelles have several morphological and biochemical similarities with particular bacteria, which led Margulis (1970) to propose that these organelles were of a bacterial origin. One of these arguments was that organelles have their own genome: this genome might be a relic of ancestral bacterial genomes. The digestion technique of Carl R. Woese confirmed that these organelles derived from bacterial ancestors.

First, Zablen *et al.* (1975); Bonen et Doolittle (1975) showed that chloroplast rRNA were more similar to cyanobacterial ones than to those from the eukaryotic nucleus. This was later confirmed by many other studies, and refined to propose that chloroplasts came from heterocyst-forming Cyanobacteria (Deusch *et al.*, 2008), *i.e.* Cyanobacteria that produce a particular cell, with a heavy cell wall, in which nitrogen is fixed into amino-acids. This suggests that all Eukarya possessing chloroplasts are younger than heterocyst-forming Cyanobacteria. As these Cyanobacteria may be between 2.45 and 2.1 billion years old (Tomitani *et al.*, 2006), chloroplast-bearing Eukarya did not appear before these dates. Rodríguez-Ezpeleta *et al.* (2005) showed that it was probably in an ancestor of Viridiplantae (plants in general), Rhodophyta (red algae), and Glaucophyta that the endosymbiosis of chloroplast took place.

Second, Bonen *et al.* (1977) showed that the power-generating mitochondria's rRNA were more similar to bacterial ones than to those from the eukaryotic nucleus. The analysis of other genes has once again confirmed these results many times, and permitted to show that mitochondria emerged from alpha-Proteobacteria (Esser *et al.*, 2004). This origin seems to make sense, as alpha-Proteobacteria contain many organisms intimately associated to Eukarya, as parasites of animals or plants for instance. One hypothesis notably proposes that alpha-Proteobacteria and a methanogenic archaea first associated, the alpha-Proteobacteria protecting the methanogen from excess oxygen thanks to its reducing waste product H_2 (Martin et Müller, 1998). Through time, the once free-living bacteria was engulfed by the methanogen, and as eons went by, the whole chimera turned into what we now know as Eukarya. This hypothesis is based on the observation that many associations are now found between methanogens and alpha-Proteobacteria; however, in phylogenies of the tree of

life no methanogenic archaea is found at the root of Eukarya. As mitochondria or mitochondrial remains have been detected in all Eukarya (Hrdy *et al.*, 2004), Eukarya are more recent than alpha-Proteobacteria. Geological studies (section 2.4) suggest that Eukarya may be at least 1.68 billion years old, thus so should be alpha-Proteobacteria.

Eukarya are shameless thieves.

Other endosymbioses have happened in the eukaryotic kingdom, hence named secondary endosymbioses. The particular propensity of Eukarya to engulf other cells may be related to their fluid cellular membrane (that contains cholesterol) and their cytoskeleton. It is interesting to note that a great part of the poor metabolic diversity shown by Eukarya was borrowed from Bacteria. From a metabolic point of view, Eukarya are annoying followers.

2.7.4 A view of the tree of life

Several studies have focused on particular parts of the tree of life, in order to improve their phylogeny. Recent works usually rely on the consideration of a large number of genes at the same time, in the hope that the average phylogenetic signal may be a good estimator for a species phylogeny. Similar approaches have been used in the three kingdoms of life.

Archaeal phylogeny

Céline Brochier, Simonetta Gribaldo and Patrick Forterre have made great efforts to improve the phylogeny of Archaea (Matte-Tailliez *et al.*, 2002; Forterre *et al.*, 2002; Brochier *et al.*, 2004, 2005b,a; Gribaldo et Brochier-Armanet, 2006; Brochier-Armanet *et al.*, 2008; Elkins *et al.*, 2008). To this end, they benefited from recently published whole genome sequences to analyse a large number of genes that are well conserved among all Archaea and therefore thought to be good markers of the species phylogeny. They argue that a well supported phylogeny is now emerging. However, it is still unclear where species such as *Cenarchaeum symbiosum* and Candidatus *Korarchaeum cryptofilum* should be placed (article 5 and Elkins *et al.* (2008)), and the diversity of Archaea still needs to be better sampled.

The ancestor of Archaea may be hyperthermophilic.

Classically, Archaea are divided in two phyla, Euryarchaeota and Crenarchaeota (Fig. 2.7). Crenarchaeota are the least well represented group, and contain Thermoproteales, Sulfolobales and Thermococcales. Euryarchaeota contain all other species, to the exception of the recently proposed Thaumarchaeota (article 5), whose position is unclear, and is figured here at the base of all Archaea, but may also be at the base of Crenarchaeota. As lots of hyperthermophilic (with optimal growth temperature above 80°C) species are found in Archaea and as they are scattered in both Euryarchaeota and Crenarchaeota, it has been as-

sumed that the ancestor of Archaea was a hyperthermophilic organism (Gribaldo et Brochier-Armanet, 2006).

Methanogenesis is a metabolism endemic to Archaea, which is practised by Methanopyrales, Methanobacteriales, Methanococcales, Methanomicrobiales and Methanosarcinales. One can parsimoniously infer that methanogenesis appeared in the last common ancestor of all these groups, which dates this node at more than 3.46 billion years ago (Fig. 2.7). If such a constraint is applied to the tree of Archaea, and if we choose for LUCA an age lower than 4 billion years, the primary radiations in Archaea seem very close to each other.

Bacterial phylogeny

A few groups of scientists have endeavoured to analyse several genes combined to propose a phylogeny of Bacteria (Battistuzzi *et al.*, 2004; Beiko *et al.*, 2005; Bern et Goldberg, 2005; Ciccarelli *et al.*, 2006; Choi et Kim, 2007; Bapteste *et al.*, 2008). I have also attempted to produce a bacterial phylogeny in article 3. Overall, all these phylogenies recover similar groupings. However, it is not clear that these proximities in the tree reveal true evolutionary relationships: several confounding factors may produce artefactual groupings.

There is a tree of Bacteria.

One confounding factor is notably found in lateral gene transfers, through which pieces of genomes can be exchanged between species. Some believe that gene transfers are so prevalent that a vertical history of the genomes, that would record speciations and not transfers, cannot be recovered (Doolittle, 1999). Several studies however suggest that a tree of the vertical history of Bacteria can be reconstructed (Daubin *et al.*, 2003; Beiko *et al.*, 2005; Choi et Kim, 2007; Galtier, 2007; Soria-Carrasco et Castresana, 2008), and have been a motivation for article 3.

What was the favourite temperature of the bacterial ancestor?

Better methods for phylogenetic reconstruction may clarify the history of life in Bacteria, that clearly distinguish between species tree and gene trees. For now, it seems that Aquificales and Thermotogales are particularly basal Bacteria (Brochier et Philippe (2002) disagree, but they relied on a single gene). As these two groups contain hyperthermophilic bacteria, it is unsure what was the favourite temperature of the ancestor of all Bacteria. Article 6 treats such questions in details.

Few geological data permit to constrain dates on nodes of the tree of Bacteria. However, alpha-Proteobacteria are necessarily older than Eukarya, as all Eukarya possess mitochondria or mitochondrial remains (see 2.7.3). This constraint implied that by 1.68 billion years ago, most bacterial phyla had already diversified. This may be consistent with Boussau *et al.* (2004) who inferred through parsi-

mony that the ancestor of all alpha-Proteobacteria may have lived in an oxygenic environment. However, much work remains to be done to clarify the bacterial tree and align it with the geological record.

Eukaryotic phylogeny

Eukaryocentrism.

The eukaryotic phylogeny has been the subject of intense research as well as intense debate, at nearly all taxonomic depths, from the phylogeny of mammalia (Li *et al.*, 1990; Graur *et al.*, 1996, 1997; Madsen *et al.*, 2001; Murphy *et al.*, 2001; Ranwez *et al.*, 2007; Wildman *et al.*, 2007), to the phylogeny of the whole kingdom (Gouy et Li, 1989; Moreira *et al.*, 2000; Philippe *et al.*, 2004; Douzery *et al.*, 2004; Rodríguez-Ezpeleta *et al.*, 2007b; Burki *et al.*, 2008), passing by the phylogeny of animals (Adoutte *et al.*, 2000; Delsuc *et al.*, 2006; Bourlat *et al.*, 2006; Marlétaz *et al.*, 2006; Dunn *et al.*, 2008), or the phylogeny of plants (Qiu *et al.*, 1999; Savolainen *et al.*, 2000; Savolainen et Chase, 2003; Davies *et al.*, 2004; Jansen *et al.*, 2007; Frohlich et Chase, 2007). Many relationships are still unresolved however, and it is for instance far from clear where the root of Eukarya should be placed. Because Eukarya are the kingdom with the greatest fossil record, people have endeavoured to analyse sequences accounting for these fossils. Fossils permit to anchor in time some nodes of a phylogeny, and thus can be used to get information about rates of evolution, but also about times. Using such data, Douzery *et al.* (2004) estimated that Eukarya diverged around 0.95-1.259 billion years ago, much more recently than the estimate derived from geology, 1.68 billion years ago (see 2.4.2). Besides technical errors, this discrepancy may come from an improper equation of cholesterol with Eukarya, or may also suggest that Eukarya as we now see them are only the top of an iceberg, the bottom of which would be now extinct.

How old are Eukarya?

Although many relationships are still unclear, the tree of life can now be depicted in broad strokes. The consideration of geological data permits to date certain points in the tree (Fig. 2.7).

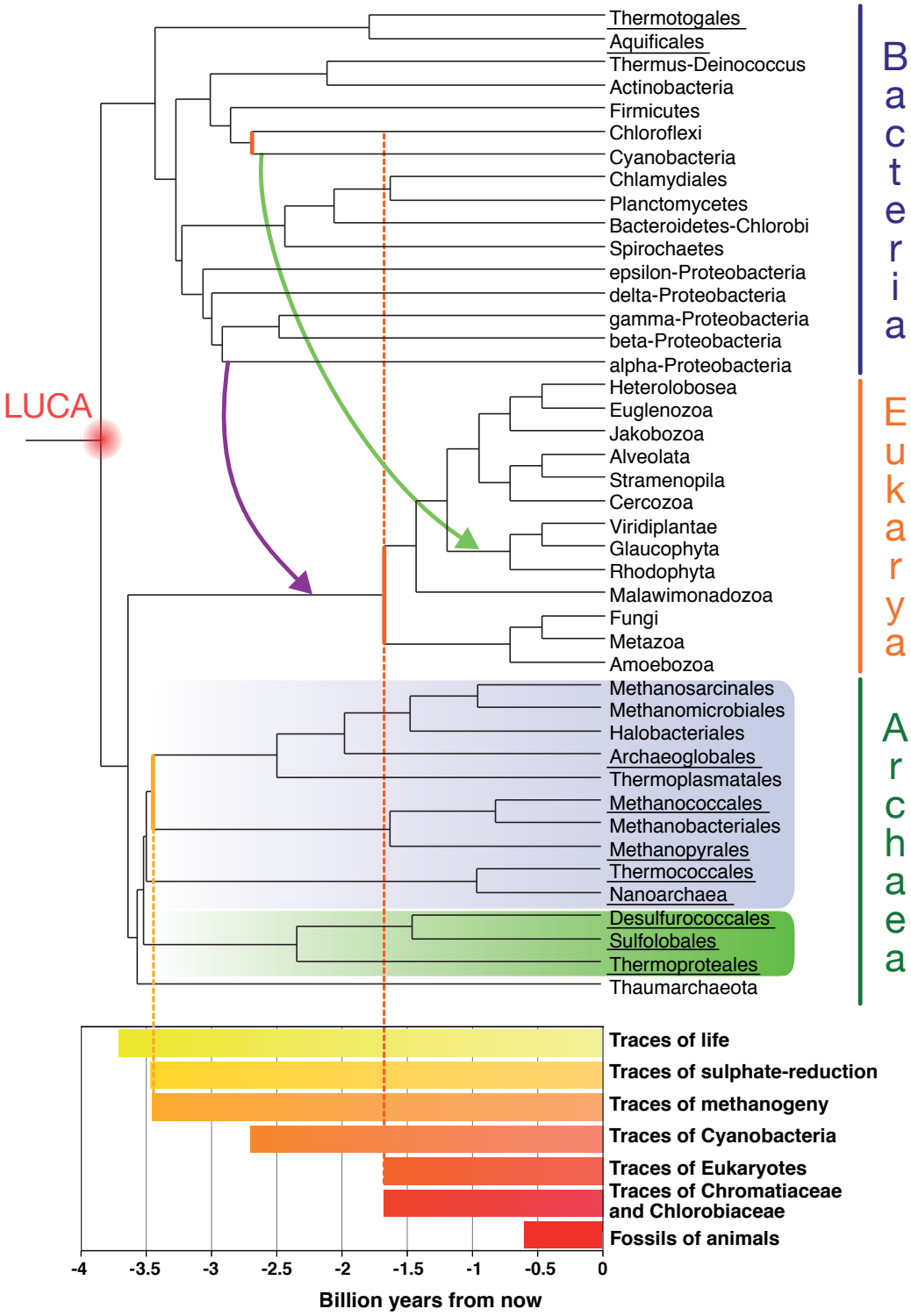


Figure 2.7: *A subjective view of the tree of life as revealed by analyses of genomes. The phylogeny of Bacteria is as in article 3, the phylogeny of Archaea has been compiled from Gribaldo et Brochier-Armanet (2006); Brochier-Armanet* et al. *(2008); Elkins* et al. *(2008), and the phylogeny of Eukarya from Rodríguez-Ezpeleta* et al. *(2007a). Some nodes of the tree have been constrained to agree with datations as obtained in section 2.4.4. Dates associated with non-constrained nodes should be ignored. Primary endosymbioses have also been represented with arrows indicating the direction of transfer: purple for the origin of mitochondria, and green for the origin of chloroplasts. Phyla harbouring hyperthermophilic organisms have been underlined. Crenarchaeota are on a green background, and Euryarchaeota on a blue background.*

Although some nodes have been dated, we still do not have a proper temporal framework to understand the evolution of life. A lot of interesting work remains to be done to improve the tree, by identifying ancient relationships, estimating the properties of extinct organisms, inferring the events that gave rise to biodiversity, and dating ancient nodes. In these purposes, the dramatic increases in the amounts of data available offer an excellent starting point. With these data in hand, better models of sequence evolution as well as better models of species tree reconstruction need to, and can, be developed. Many people are working on such projects right now, and the field is developing at an exciting pace. My own thesis has revolved around such models of genome and sequence evolution.

2.8 Organisation of the manuscript

Pattern, process, and early evolution.

My thesis work has been in line with efforts to improve our knowledge of the early evolution of life. I have tried to improve techniques of phylogenetic reconstructions, and have used some of these techniques to propose answers to particular evolutionary problems.
This manuscript contains eight articles I have contributed to. I present them in a non-chronological order, and start by an article that presents examples of difficulties that are met when one attempts to phylogenetically place a species. Then I present articles where these difficulties have been addressed by the development of new methods, and examples of application of these new methods.

- The first article (3) attempts to clarify the phylogenetic position of Aquificales, a group of bacteria living in hot environments. Their phylogenetic placement is important to our understanding of the evolution of tolerance to high temperatures, but has been difficult to estimate, as their genome seems to contain lots of genes coming from other organisms, and has also evolved in a peculiar manner under the influence of extreme life conditions (it has developed a *compositional bias*). Although I could not get rid of all potential artifacts, several clues suggest that they may be related to Thermotogales, another lineage of heat-loving organisms. This study showed me some of the major difficulties that one has to face when trying to do "deep" phylogenetics, and convinced me that better models of sequence evolution should be developed. Notably, they motivated me to work on models robust to compositional bias (articles 4 and 7), recombination (article 8), and lateral gene transfers (article 10).

- The second article (4), although the product of an earlier work, can be seen as a very partial answer to the problems that the first article raised, as it tackles the issue of compositional bias. In this article, we have shown that models of sequence evolution more robust to compositional biases than

commonly used ones could be used as easily, by looking carefully at how the likelihood of a tree is computed. This idea was exemplified through the development of a piece of software, nhPhyML, that we showed was more robust to compositional bias than other common methods.

- This software was then used to try and place a particular organism, the archaea *Cenarchaeum symbiosum* into the tree of life. We propose that *C.symbiosum* may represent a third archaeal phylum, in addition to Euryarchaeota and Crenarchaeota, as it branches far from the other archaea, and as its gene content is distinct from both other phyla (article 5). My little contribution did not bring very much, as nhPhyML did not deliver a firm answer as to *Cenarchaeum*'s relationships. I believe some improvements could be applied to the capabilities of nhPhyML to cope with high numbers of sequences while correctly exploring the space of tree topologies.

- nhPhyML may not be great at exploring the space of tree topologies, it can however accurately estimate the content in bases G and C in ancestral sequences, as I had shown in the second article. In rRNA, this G+C content is correlated to the host organism's optimal growth temperature, so that estimating one permits to infer the other. Using this software as well as a Bayesian software developed by co-authors of this article (6), we propose that LUCA was much less a heat-loving organism than its two descendants. This surprising pattern is in agreement with previously published hypotheses, and suggests ways through which geology and evolutionary biology may illuminate each other.

- The previous article benefited from a Bayesian software from our co-authors, but it also showed me that much progress remains to be done to routinely use models of sequence evolution that are robust to compositional bias, as their program took weeks to run on a fixed topology and for 30 sequences. I believe that the work presented in this article (7) may be a step in the right direction, as it should help phylogeneticists easily test new ideas.

- Article (8) deals with another problem identified in the first article, that of recombination, by which a gene is the product of two or more different evolutionary histories. I propose two models to reconstruct these evolutionary histories, and test them.

- The last difficulty identified in the first article is the fact that gene trees can differ from the species tree. In this seventh article (9), I present a model that separately infers a species tree from gene trees.

- This last article (10) also works as a conclusion to this manuscript. In the course of my thesis, I have come to understand that probabilistic models could provide great insight into the history of life by using the information

contained in genomes. In this review, we present some recent progress that has been made in models of evolution, and propose perspectives that we believe are very promising.

- As appendices, I have added two articles I contributed to before my thesis. One focuses on interpreting phylogenies of genes present in Chordates, and the other attempts to reconstruct gene content evolution in alpha-Proteobacteria.

3

Phylogeny is not easy

This first article attempts to clarify the phylogenetic position of a particular group of Bacteria, Aquificales. Meanwhile, several problems related to molecular phylogeny are treated.

First, genomes from organisms living in similar environments can have converged to similar characteristics. In the present case notably, Aquificales may be artefactually grouped with Thermotogales because they share similar sequence compositions. This compositional bias can mislead phylogenetic reconstruction. Although we have tried in this article to diminish its impact, a better answer may come from better models of evolution.

Second, lateral (or horizontal) gene transfer (LGT) can considerably alter phylogeny reconstruction, as in its presence, gene trees can differ from species trees. Once again, in this article we have tried to do our best to infer a species tree despite LGT, but better ways to do so could be found in new models of evolution.

In the end, our results suggest that Aquificales may be more related to Thermotogales than to other bacteria. They also call for better models of evolution that could cope both with compositional biases and gene transfer.

This article has been accepted for publication in *BMC Evolutionary Biology*.

Accompanying Supplementary Materials can be found at the following addresses:

`http://biomserv.univ-lyon1.fr/~boussau/Article1/AdditionalFile1.xls`

`http://biomserv.univ-lyon1.fr/~boussau/Article1/AdditionalFile2.pdf`

BMC Evolutionary Biology

Research article

Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria

Bastien Boussau*, Laurent Guéguen and Manolo Gouy

Address: Université de Lyon; Université Lyon 1; CNRS; INRIA; Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France

Email: Bastien Boussau* - boussau@biomserv.univ-lyon1.fr; Laurent Guéguen - gueguen@biomserv.univ-lyon1.fr; Manolo Gouy - mgouy@biomserv.univ-lyon1.fr

* Corresponding author

Published: 3 October 2008

BMC Evolutionary Biology 2008, **8**:272 doi:10.1186/1471-2148-8-272

Received: 14 May 2008
Accepted: 3 October 2008

This article is available from: http://www.biomedcentral.com/1471-2148/8/272

Abstract

Background: Despite a large agreement between ribosomal RNA and concatenated protein phylogenies, the phylogenetic tree of the bacterial domain remains uncertain in its deepest nodes. For instance, the position of the hyperthermophilic Aquificales is debated, as their commonly observed position close to Thermotogales may proceed from horizontal gene transfers, long branch attraction or compositional biases, and may not represent vertical descent. Indeed, another view, based on the analysis of rare genomic changes, places Aquificales close to epsilon-Proteobacteria.

Results: To get a whole genome view of *Aquifex* relationships, all trees containing sequences from *Aquifex* in the HOGENOM database were surveyed. This study revealed that *Aquifex* is most often found as a neighbour to Thermotogales. Moreover, informational genes, which appeared to be less often transferred to the *Aquifex* lineage than non-informational genes, most often placed Aquificales close to Thermotogales. To ensure these results did not come from long branch attraction or compositional artefacts, a subset of carefully chosen proteins from a wide range of bacterial species was selected for further scrutiny. Among these genes, two phylogenetic hypotheses were found to be significantly more likely than the others: the most likely hypothesis placed Aquificales as a neighbour to Thermotogales, and the second one with epsilon-Proteobacteria. We characterized the genes that supported each of these two hypotheses, and found that differences in rates of evolution or in amino-acid compositions could not explain the presence of two incongruent phylogenetic signals in the alignment. Instead, evidence for a large Horizontal Gene Transfer between Aquificales and epsilon-Proteobacteria was found.

Conclusion: Methods based on concatenated informational proteins and methods based on character cladistics led to different conclusions regarding the position of Aquificales because this lineage has undergone many horizontal gene transfers. However, if a tree of vertical descent can be reconstructed for Bacteria, our results suggest Aquificales should be placed close to Thermotogales.

Background

In the study of evolution, as in any scientific endeavour, progress relies on the comparison of hypotheses with respect to how well these succeed in accounting for a range of observed data. In phylogenetics, a given tree, a hypothesis, is confronted with trees inferred using other data; resulting incongruences are then explained by a methodological artefact, or the inability of a single tree to properly depict the evolution of the biological entities under consideration. The large agreement between the ribosomal RNA (rRNA) bacterial phylogeny and phylogenies built from a concatenated set of protein sequences was therefore a strong piece of evidence that the tree of life could be solved [1]. For instance, protein phylogenies confirmed the monophyly of most rRNA-defined bacterial phyla. Similarly, Aquificales are found close to Thermotogales both in trees built from rRNA and from concatenated proteins. However, the position of the Aquificales clade within the phylogeny of Bacteria has often been questioned on the ground of single gene phylogenies, phylogenies built from gene or domain content [2], and supposedly rare genomic changes such as insertions-deletions [3-8]. Strikingly, many of these analyses are congruent with each other and suggest that Aquificales might be more closely related to Proteobacteria than to Thermotogales. This new view has been adopted in recent scenarios that explain the whole evolution of life on earth [9], so it is important to our understanding of bacterial evolution that the puzzling phylogenetic problem of the position of Aquificales within the bacterial phylogeny gets solved.

Species phylogenies built from the comparison of gene sequences suffer from two major limitations: on one side the true gene trees may differ from the species trees, and on the other side, the signal contained in the gene sequences might be too weak or too complex to be correctly interpreted by bioinformatics methods. Gene trees will differ from species trees in cases of hidden paralogy, closely spaced cladogenesis events or horizontal gene transfers (HGT). This last phenomenon is particularly relevant to the present study, as gene transfers are frequent among prokaryotes. Phylogeneticists therefore often only consider informational genes, involved in the processes of transcription, translation and replication, which appear to be less prone to HGTs over broad distances than other genes, named operational [10]. The second limitation, that of a phylogenetic signal so blurred or buried that tree reconstruction methods fail to recover the true tree, may come from a saturated history of mutations (long branch attraction, [11,12]) or compositional biases [13,14]. Both pitfalls are likely to affect genes used to reconstruct the bacterial phylogeny, because Bacteria possibly date as far back as 3.5 billion years ago [15], and because they display a great diversity in their genomic characteristics as well as in their ecological niches. More specifically, Aquificales may be placed close to Thermotogales not because they last diverged from them, but because they share a common ecological niche, *i.e.* they are both hyperthermophilic, which led both their rRNA [16] and their protein sequences [17] to adapt to high temperatures. Sequence similarities between these two clades would therefore be the result of convergences due to identical selective pressures, not the result of common descent. Consequently, recovering the bacterial species tree and clarifying the relations between hyperthermophilic organisms from comparison of gene sequences is a difficult task, and has led several authors to search for more reliable informative characters.

Such characters are cell-structural features, or of a genomic nature: "rare genomic changes" [18], such as gene fusion/fission or insertion-deletions (indels), and gene or domain presence/absence. The main assumption concerning all these characters is that they are nearly immune to convergence: to be informative, a given character, morphological or genetic, should only arise once. To our knowledge, this assumption has never been thoroughly tested. The genomic characters further depend on the identification of orthologous genes in different genomes, and consequently are subject to the pitfall of horizontal gene transfers. Here again, this weakness is of particular interest to our study, since both Aquificales and Thermotogales seem to be particularly prone to exchanging genes with other bacterial species [19,20].

Therefore it appears that both approaches – sequence phylogenies and character cladistics – are potentially hindered by defaults whose magnitude is sufficient to question their conclusions. As in the case of the phylogenetic position of Aquificales their conclusions diverge, a detailed study might clarify which approach has suffered most from its drawbacks.

In this report, we used the HOGENOM [21] database to survey the phylogenetic neighbourhood of *Aquifex*. This database contains families of homologous genes from complete genome sequences with associated sequence alignments and maximum likelihood phylogenetic trees. The automatic survey of all trees containing sequences from *Aquifex* in the HOGENOM database reveals that *Aquifex* is most often found as a neighbour to Thermotogales. When genes are separated into informational and non-informational genes we find that genes from the former category seem to be less transferred than non-informational ones. To this end, neighbour clades for each gene from *Aquifex* were counted, separately for informational genes and for operational genes, yielding two distributions. Then for each of the two distributions, Shannon's index of diversity was computed [22]. This

index measures whether the genes are evenly distributed among all possible neighbourhoods or whether a specific vicinity dominates. We find that the index value is significantly different between the two distributions: among informational genes, one neighbourhood, between Aquificales and Thermotogales, tends to dominate the distribution much more than in operational genes. This shows that there is one dominating phylogenetic signal among informational genes, and much less among operational genes, which is consistent with the idea that operational genes experience more frequent HGT events than informational genes.

To study the impact of saturation and compositional heterogeneity on the position of Aquificales, we concatenated a large dataset of putatively orthologous proteins from a wide range of bacterial species (Additional file 1). A phylogenetic tree was built, and then taken as a reference to test for the position of Aquificales: Aquificales were first removed from the tree, and then re-introduced in the topology in all possible positions. Site likelihoods were computed for all these positions, which allowed for the identification of sites favouring a given topology. Two phylogenetic hypotheses were found to be significantly more likely than the others: the most likely hypothesis placed Aquificales as a neighbour to Thermotogales, and the second one placed Aquificales with epsilon-Proteobacteria. We characterized the genes that supported each hypothesis, and found that differences in rates of evolution or in amino-acid compositions could not explain the presence of two dominating phylogenetic signals in the alignment. However, evidence for a large Horizontal Gene Transfer between Aquificales and epsilon-Proteobacteria was found. These findings suffice to explain why methods based on concatenated informational proteins and methods based on character cladistics led to different conclusions, and suggest that the vertical signal in the genomes of Aquificales, *i.e*. the portion of the genome most likely to have been inherited through descent and not through HGT, relates them to Thermotogales.

Results and discussion

A whole genome view of Aquifex *relationships*

For each gene tree containing sequences from *Aquifex aeolicus* in the HOGENOM database, the identity of the group of sequences neighbouring *Aquifex* was recorded. This gave counts of *Aquifex* genes found close to Thermotogales, Firmicutes, epsilon-Proteobacteria, among others. Cases where *Aquifex* genes were found close to a non-monophyletic group of species were discarded, which left 578 gene trees. Among these, *Thermotoga* is found as *Aquifex's* closest neighbour 98 times, epsilon-Proteobacteria are found 44 times, delta-Proteobacteria 84 times, Firmicutes 71, Thermus-Deinococcus 39, Euryarchaeota 74 (see Fig. 1). In view of such a distribution, it is difficult to argue in favour of any particular relationship: Horizontal Gene Transfers appear so pervasive that no signal emerges as clearly dominant. However, HGTs may not affect all types of genes with similar frequencies. It has been proposed that genes that are related to the universal processes of transcription, translation and replication and known as "informational genes" may be less transferred than "operational genes", involved in metabolism for instance [10].

We therefore separated HOGENOM protein families into informational and non-informational gene families. Fig. 1a shows that among informational genes, the genes placing *Aquifex* close to *Thermotoga* (32 genes) are twice more numerous than the genes favouring the second best alternative hypothesis, *i.e*. the vicinity of Firmicutes (15 genes). On the contrary, among operational genes (Fig. 1b), differences between various hypotheses are much narrower: *Thermotoga* is *Aquifex's* neighbour in only two more cases than delta-Proteobacteria, 11 more cases than Firmicutes, and 13 more cases than Euryarchaeota. To quantify this comparison, Shannon's index of diversity was measured for both sets of genes. This index measures how evenly distributed observations are among categories [22]: the higher the index, the more even the distribution; conversely, the lower the index, the more a few categories dominate. Shannon index values were 2.07 for informational genes, and 2.49 for operational genes (significantly different according to a t-test, p-value < 0.001; a Pearson χ^2 test between the two distributions is also significant, p-value < 10^{-20}), which means that operational genes are significantly more evenly distributed among the various neighbour groups than informational genes. The distributions depicted in Figs 1a and 1b result from a mixture of lack of phylogenetic resolution at the single-gene level and of HGT events. But the difference between them strongly suggests that operational genes have been horizontally transferred more often than informational genes, which is consistent with the fact that Euryarchaeota are almost never found as neighbour to *Aquifex* in informational genes (2%), but often found in operational genes (11%). Interestingly, for both sets of genes, epsilon-Proteobacteria are not one of the most frequent *Aquifex* neighbours, as they are less frequent than *Thermotoga*, Firmicutes, and delta-Proteobacteria. For operational genes, they are even less frequent than Euryarchaeota. These results thus do not support the hypothesis that Aquificales are epsilon-Proteobacteria [4]. However, if all Proteobacteria are to be counted as a single clade, the vicinity of *Aquifex* with Proteobacteria becomes a high-scoring hypothesis: *Aquifex* is most closely related to a Proteobacterium with 18 informational genes and 76 non-informational genes. According to operational genes, if anything, *Aquifex* would be a Proteobacterium, as almost twice more genes place it with Proteobacteria than with *Thermotoga* (76 for Proteobacteria against 41 for *Thermotoga*); accord-

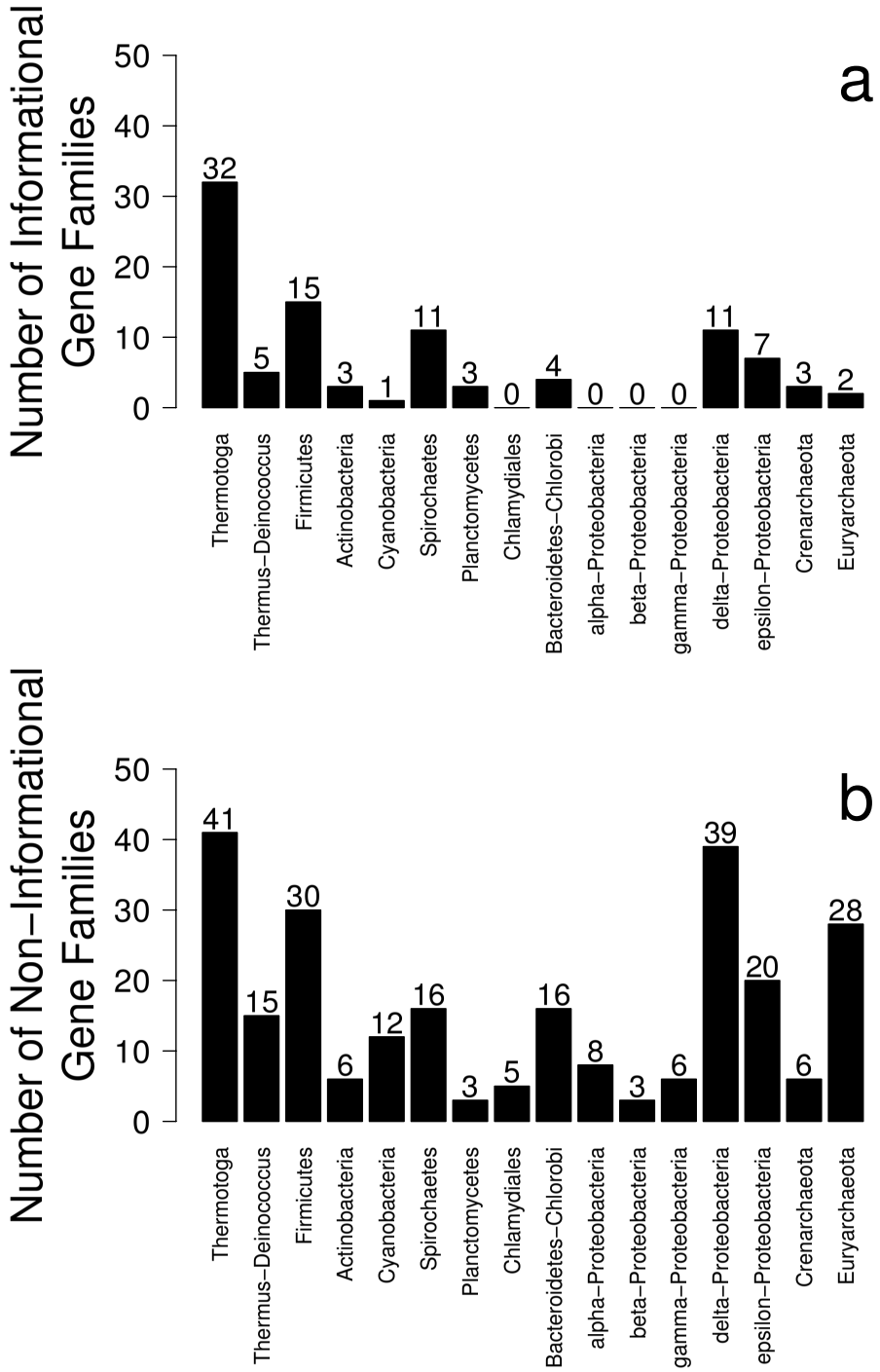


Figure 1
Phylogenetic relationships of *Aquifex* genes according to the HOGENOM database. a: Informational genes. b: Non-informational genes.

ing to informational genes, *Aquifex* is close to *Thermotoga*, as almost twice more genes place it with *Thermotoga* than with Proteobacteria (18 for Proteobacteria against 32 for *Thermotoga*). However, considering all Proteobacteria as a single clade artificially groups a variety of different histories under the same hypothesis. It is thus more likely that the high frequency of close relationships between *Aquifex* and *Thermotoga* among informational genes reflects vertical descent, and that the scattered distribution of *Aquifex* closest homologs among operational genes results from frequent horizontal transfers to or from the *Aquifex* lineage.

Furthermore, this whole genome analysis may suffer from compositional biases or long branch attraction. Consequently, a subset of carefully chosen genes was concatenated and used to assess the importance of potential artefacts: first a tree of the Bacteria was built, and then, using this tree as a scaffold, the influence of saturation and compositional biases on the position of Aquificales was estimated.

Bacterial phylogeny obtained from a concatenated set of putatively orthologous genes

Fifty-six genes that were nearly universal in Bacteria and present as single copy in most genomes were concatenated (see Methods). Genes that showed a transfer between Bacteria and Archaea had previously been discarded because a gene showing evidence of a transfer between very distantly related organisms might be especially prone to be transferred among species of the same domain. Some of the 56 remaining genes may still have undergone a transfer, and concatenating them may lead to spurious results. Usually, transferred genes are discarded before gene concatenation [23,24]. Here, we first checked for possible tree building biases resulting from composition or evolutionary rate effects before proceeding to an analysis designed to specifically identify genes that may have been transferred between *Aquificales* and other species. PhyML was used to build a starting phylogeny based on the concatenated protein alignments, using the JTT model and a gamma law discretized in four classes to account for variation in the evolutionary rates. The discretized gamma law [25] is widely used because of its mathematical convenience, not as a precise model of the evolutionary rates of protein sequences. Therefore it is expected that some sites are not properly modelled when this approximation is made. To estimate how sites were modelled by the discretized gamma law, we plotted the distribution of expected relative evolutionary rates across sites (Fig. 2) as found by BppML. This distribution shows four peaks, each corresponding to the rate of a particular class. The two largest peaks are at the limits of the distribution: they comprise both sites whose rate is properly approximated by one of the two extreme evolutionary rates, but also sites whose rate would be smaller or larger, if the discretized gamma law was able to provide a convenient rate. For instance, the leftmost peak contains sites properly modelled by a relative rate of ~0.2, but also sites evolving more slowly, such as constant sites. *Per se*, improperly modelling constant sites probably does not lead to biased phylogenetic estimations; however underestimating the evolutionary rate of some fast-evolving sites (and this may be a by-product of improper modelling of constant sites) will lead to an underestimation of the convergence probability. Such misspecified modelling is therefore a potential cause for long branch attraction, as underlined in another context [26]. We consequently decided to conservatively discard sites whose evolutionary rate was above the arbitrary threshold of 2.2 (red line, Fig.2), in the hope of reducing risks of reconstruction artefacts. The resulting alignment contains 10,000 sites, and has been submitted to an additional reconstruction through PhyML, with a bootstrap analysis based upon 200 replicates.

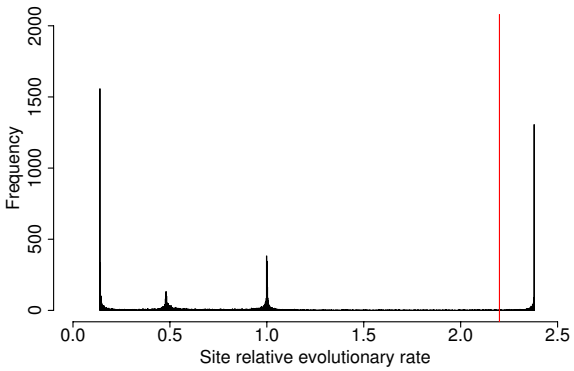


Figure 2
Distribution of the site relative evolutionary rates. Rates were estimated using a 4 class discretized gamma distribution. The 4 peaks correspond to the rates associated to each class. The vertical red line corresponds to the threshold above which sites have been discarded due to their high evolutionary rate.

Our tree comprises 94 bacterial species, spanning as exhaustively as currently possible the diversity of Bacteria (Fig. 3). The resulting topology is in good agreement with rRNA trees [27], recently published concatenated-protein phylogenies [28,29], as well as supertree phylogenies [30]. In particular, we do recover the clade named "Terrabacteria" by Battistuzzi and co-workers, as well as the clade named Gracilicutes by Cavalier-Smith [7], separated with a high bootstrap support (BS 94%). It is interesting to note that these three recent bacterial phylogenies all

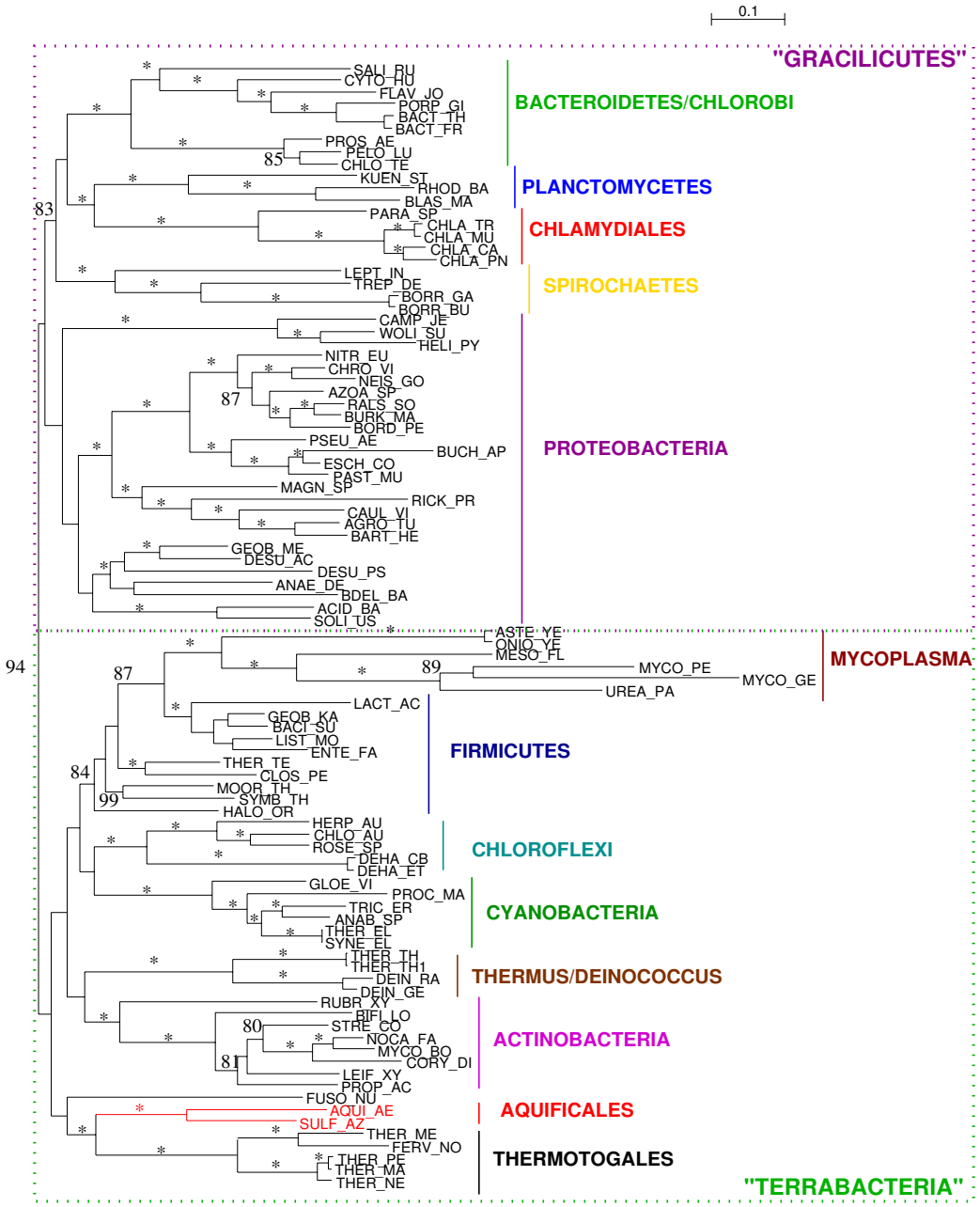


Figure 3

Unrooted phylogenetic tree of Bacteria. This tree was obtained after discarding all sites with evolutionary rate predicted to be above 2.2. Stars indicate branches with 100% bootstrap support (200 replicates). Bootstrap supports between 80% and 100% are shown, bootstraps below 80% have been removed for clarity. Aquificales are represented in bright red. Names of major groups are according to the NCBI taxonomy. Gracilicutes and Terrabacteria, two recently proposed superclades, are shown as dashed frames, and their names are between quotation marks to mark their unconsensual status.

recover these two clades, which suggests that the global picture of bacterial evolution might be slowly unveiling. The "PVC supergroup" (Planctomyces-Verrucomicrobia-Chlamydiales, [31]) seems to find a confirmation in our phylogeny where Planctomycetes and Chlamydiales are grouped with 100% BS. Many similarities are also found

with the phylogeny proposed by Ciccarelli and co-workers [32], or the supertree obtained by Beiko, Harlow and Ragan [33], such as the monophyly of Proteobacteria, and the grouping of Aquificales with Thermotogales.

However, many deep nodes do not obtain high bootstrap supports. Two avenues might help fully resolve the bacterial phylogeny: further increase the number of phylogenetic markers, and improve the interpretation of the phylogenetic signal through the development of new models of evolution. Such models would ideally be able to deal with compositional heterogeneity, and would safely handle saturation. As there is no efficient program with these properties, we have chosen to filter out saturated sites to try and diminish compositional heterogeneity.

We have already attempted to remove the most saturated sites. To assess the impact of compositional heterogeneity, we performed Bowker's tests for symmetry in the evolutionary process on the whole alignment [34,35]. Bowker's test relies on the comparison of two sequences against each other, therefore 94*93/2 = 4371 tests can be done on our alignment. Among these 4371 tests, 3826 reject symmetry at the 5% level: though we have made no effort to alleviate the multiple tests problem, compositional heterogeneity might be an important issue for the reconstruction of bacterial phylogeny. Species that show the most biased amino-acid usage, *i.e*. that fail the highest numbers of Bowker's tests, include first AT-rich species (*Buchnera aphidicola*, *Borrelia burgdorferi*), then GC-rich species (*Thermus Thermophilus*) and finally hyperthermophilic species (data not shown). This is in agreement with results based on a multivariate analysis of proteome composition [36], where the GC content of the genome was found to be the major factor influencing amino-acid composition, before thermophily.

To try and limit the influence of compositional bias, we recoded the concatenated protein alignment in 4 states based on the physico-chemical properties of the amino-acids [37]. Such a recoding is expected to reduce the risk of long branch attraction artefact as well as compositional bias by decreasing the number of homoplasies. Accordingly, after the recoding, 2818 tests reject symmetry: the recoding seems to have diminished compositional bias at least in 1008 cases, but clearly has not permitted to fully erase heterogeneity. The tree we obtain on the recoded alignment (Fig. 4) is very similar to the previous tree (Fig. 3), with Gracilicutes separated from Terrabacteria (BS 76%). Interestingly, Aquificales are still found as a sister group of Thermotogales with a high bootstrap support (96%), and Thermus-Deinococcus also clusters with these hyperthermophilic organisms, although the bootstrap support is negligible (36%). The grouping of the photosynthetic lineages Chloroflexi and Cyanobacteria gains support through the recoding, with a BS of 85% on the recoded alignment against 77% on the original alignment. So does the clustering of these two photosynthetic lineages with another lineage that contains photosynthetic organisms, the Firmicutes: from 63% on the original alignment, the BS increases to 73% with the recoded alignment. The grouping of these three photosynthetic lineages appears as an appealing hypothesis, but certainly requires further inquiry, especially since horizontal gene transfers are thought to have been part of the evolution of photosynthesis [38]. Strikingly, Spirochaetes were found to group with Chlamydiales, Planctomycetes and Bacteroidetes/Chlorobi with a high bootstrap support (83%) on the original alignment, but grouped with epsilon-Proteobacteria on the recoded alignment (bootstrap support: 18%), which shows that recoding can impact tree reconstruction. Overall, the average bootstrap support is 87.1%, not significantly lower than the average support for the original alignment (90.3%, p-value = 0.065 with a Student paired t-test, p-value = 0.154 with a Wilcoxon signed rank test). This supports the conclusion of Susko and Roger [39] that recoding does not lead to a substantial loss of information.

As the trees obtained on the recoded and original alignments are in strong agreement, we conclude that we obtain a fairly robust Bacterial tree, and that the clustering of Aquificales and Thermotogales does not seem due to saturation or compositional artefacts. However, since more than 50% of Bowker's tests reject symmetry on the recoded alignment, considerable compositional heterogeneity has escaped the 4-state recoding, and this analysis cannot entirely rule out the hypothesis that Aquificales and Thermotogales are attracted by compositional biases. Nonetheless, the addition to the concatenated alignment of sequences from two free-living epsilon-Proteobacteria, *Sulfurovum* NBC37-1 and thermophilic *Nitratiruptor* SB155-2 [40], does not affect this grouping either (see additional file 2). Thus the Aquificales-Thermotogales grouping does not seem to result from compositional biases.

Does the Thermotogales-Aquificales cluster come from a reconstruction artefact?

The topology that is found without Aquificales using PhyML with the same parameters is perfectly congruent with the tree obtained with Aquificales. Taking therefore as reference the tree without Aquificales, we tested all possible positions for this group in the bacterial tree. The most likely position was as found by the tree search heuristics, with Thermotogales. The second most likely position was very close, at the base of a clade comprising both Thermotogales and Fusobacterium, and the third most likely position was with epsilon-Proteobacteria, the only

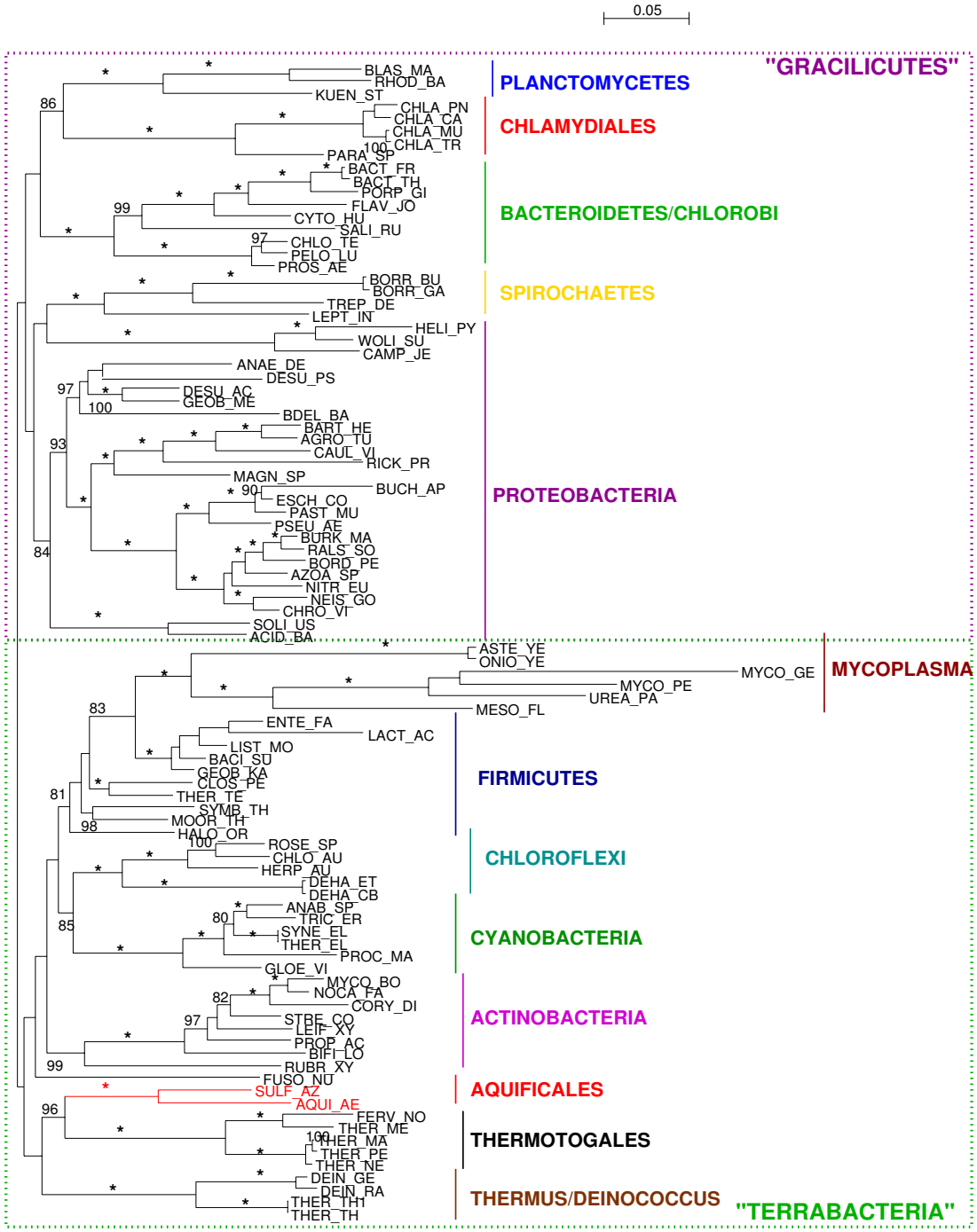


Figure 4
Unrooted phylogenetic tree obtained from 56 genes of Bacteria based on the recoded alignment. Labels as in Fig. 3.

placement not rejected at the 5% level according to an AU test [41] as implemented in Consel [42] (p-value = 0.062). Because the AU test is based on a multiscale RELL bootstrap procedure, the fact that the second most likely hypothesis is rejected by the AU test at 5% while the third is not suggests that sites of high likelihood scores are the same in the two first hypotheses, but are different from the sites of high likelihood scores in the third hypothesis. Consequently two contrasting signals can be found in the data, coming from different sites in the alignment, that support the two currently prevailing phylogenetic hypotheses for Aquificales, one based on rRNA trees, and the other heralded by Cavalier-Smith [4]. We decided to further analyse the nature of the signal that favoured each of these two placements, through a gene-wise analysis.

We built phylogenetic trees for each of our 56 genes with PhyML. Among these 56 trees, 11 place Aquificales close to Thermotogales (T genes), and only two place Aquificales close to epsilon-Proteobacteria (E genes). We compared these two sets of genes, with respect to rates of evolution and amino-acid composition, to see whether one signal is the result of a long branch attraction or of a compositional bias.

First, we computed the sum of the branch lengths for each tree in our two datasets, and computed an average branch length for each dataset. The average branch length was 0.163 for T genes, and 0.131 for E genes, which is not significantly different according to an unpaired t-test (p-value: 0.145). The discrepancy between the two datasets does not seem to be explainable by a long branch attraction artefact.

Second, the position close to Thermotogales might be favoured because of convergences instead of common descent: as written above, both Thermotogales and Aquificales are hyperthermophilic organisms, so their sequences are subject to partly similar selective pressures. Through the analysis of many completely sequenced genomes, Zeldovich and co-workers [17] have found a positive correlation between the proteome content in amino acids IVYWREL and the organism optimal growth temperature. As hyperthermophilic bacteria and archaea are not monophyletic, this suggests that there exists a selective pressure to increase the IVYWREL content in organisms that thrive best at high temperatures. If we find a higher proportion of the amino-acids IVYWREL in the Aquificales sequences for T genes than for E genes, this would imply that composition biases could be at the origin of the signal favouring the Thermotogales placement. We find that T genes in *Aquifex aeolicus* and *Sulfurihydrogenibium azorense* contain 45,4% of IVYWREL amino-acids, against 44.4% for E genes. As the difference is not significant (χ^2 test, p-value = 0.61), there is no evidence that the T signal is coming from compositional artefacts.

Consequently it appears that neither the signal favouring a close relationship between Aquificales and epsilon-Proteobacteria nor the signal favouring a close relationship between Aquificales and Thermotogales seem induced by a reconstruction artefact, namely long branch attraction or compositional convergence. Similarly, this suggests that the trees placing Aquificales close to Thermotogales in the whole genome study may not come from long branch attraction or compositional artefacts. Therefore, incongruences found between the T and E groups of genes probably unveil different gene histories: at least one of these two prevailing signals comes from HGTs.

Detection of Horizontal Gene Transfers in the concatenate

We used the 181 possible Aquificales positions whose likelihoods had been computed earlier to search for evidence of HGTs affecting Aquificales genes. Because the taxonomic sampling was as exhaustive as currently possible, and because all possible positions for Aquificales among Bacteria have been tried, it is expected that few HGTs affecting Aquificales might escape this screening.

Naturally, some genes from other Bacteria present in the dataset also underwent transfers that will not be detected using our approach. But neglecting such transfers should not affect our results, since the focus of this study is the position of Aquificales,.

The top curve of Fig. 5 shows the cumulative sum of the log-likelihood differences between the tree in which Aquificales are close to epsilon-Proteobacteria and the tree in which Aquificales are close to Thermotogales. If asked to divide this curve, one would probably cut it in two parts, the first one decreasing, and the second one increasing. This would plead for two signals, first one in favour of the Thermotogales position, and then one in favour of the epsilon-Proteobacterial position. However, this division would be based on the comparison of only two trees, whereas 181 different positions should be compared.

We used the Maximum Predictive Partitioning (MPP) algorithm to find what are the two prevailing signals in the alignment among all 181 compared positions [43]. This algorithm identifies the best way of dividing the data in two parts and assigning each to a specific tree position. The results are displayed in the bottom panel of Fig. 5. The MPP algorithm divides the alignment very close to the site in which the curve changes from descending to ascending trends. The most likely positions affected to each of the two parts, among all 181 possible positions, are first the

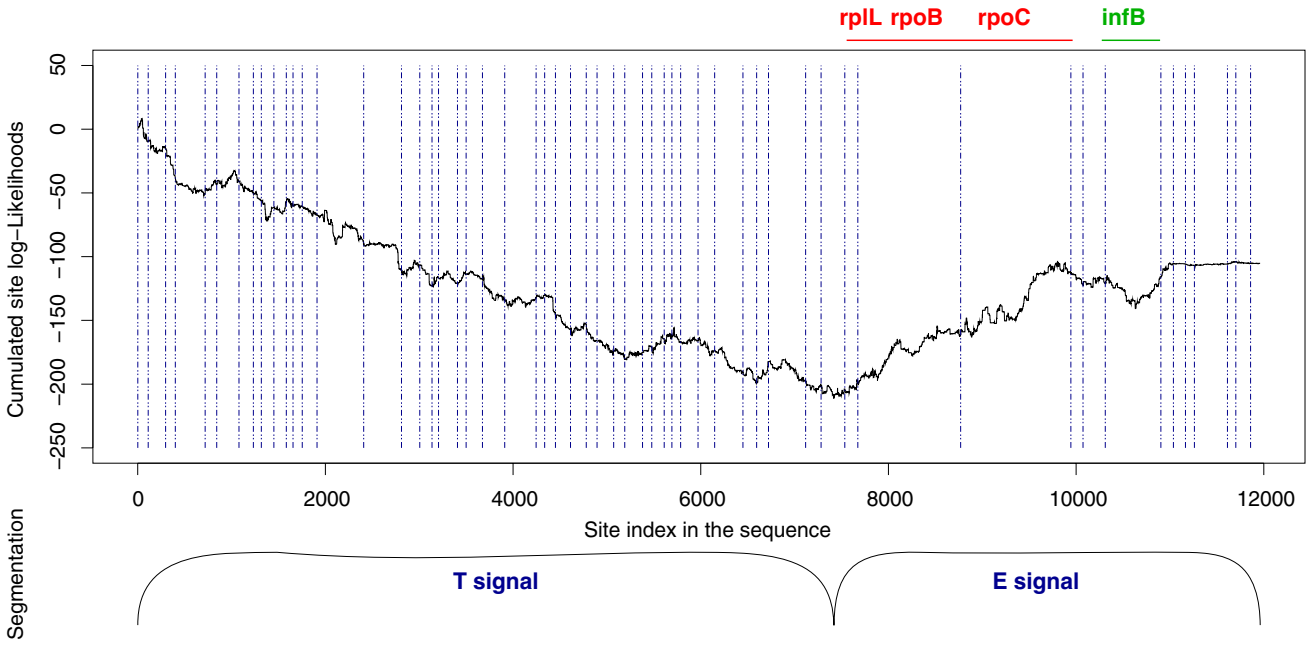


Figure 5

Comparison between site likelihoods when Aquificales are placed close to Epsilon-proteobacteria and when they are placed with Thermotogales. Upper panel: summed differences between site log-likelihoods obtained when Aquificales are placed with epsilon-Proteobacteria and when they are placed with Thermotogales. A descending trend means that a consecutive series of sites favours the Thermotogales position (T signal), whereas an ascending trend means that a series of sites favours the epsilon-proteobacterial position (E signal). Genes have been ordered according to their position along the *Aquifex* genome. Dashed blue lines represent gene boundaries. The red interval represents the genes which appear to contain most of the E signal. The green interval represents gene infB, in which the curve first decreases and then increases. Lower panel: result obtained by the Maximum Predictive Partitioning algorithm when asked to find the most likely partition of the sites in two segments. The *a posteriori* most likely model for the first segment is the tree in which Aquificales are sister group to Thermotogales, and the second segment is best fitted by the tree in which Aquificales are sister group to epsilon-Proteobacteria.

tree in which Aquificales are close to Thermotogales, and second the tree in which Aquificales are close to epsilon-Proteobacteria. Therefore, the two dominant signals in the alignment are T and E signals. Furthermore, the sequence concatenate was built following the gene order in the *Aquifex aeolicus* genome. Consequently, the fact that series of consecutive sites support the same phylogenetic position for *Aquifex* means that whole genes plead for each hypothesis.

The issue now is to decide which of these two dominant signals is most likely HGT, and which has the highest chance of coming from vertical inheritance. One can rely on the *Aquifex aeolicus* genomic map to find the solution: if a hypothesis is favoured by an isolated island that concentrates a few genes, it is likely to be the signature of a large horizontal transfer affecting a unique region of the genome. Contrary to the T signal, the signal that favours a close relationship between Aquificales and epsilon-Proteobacteria is limited to a few clustered genes, mainly consisting of the rplL-rpoB-rpoC operon (characterized in *E. coli*, [44,45]), which seems conserved in most bacterial genomes. This clustering strongly suggests that the epsilon-proteobacterial signal comes from horizontally transferred genes, through a single transfer of the whole rplL-rpoB-rpoC operon, from epsilon-Proteobacteria to Aquificales. Indeed, if only these three genes are concatenated and submitted to phylogenetic analysis, Aquificales are found clustered with epsilon-Proteobacteria with a fairly high bootstrap support (79%, Fig. 6). As these transferred genes are large, they contribute a substantial amount of signal in the complete concatenate. This large transfer appears unexpected, since it concerns informational genes, involved in translation (rplL) and transcription (rpoB-rpoC), but it has already been suggested by Iyer, Koonin and Aravind [46]; the alternative hypothesis of the E signal being the real phylogenetic signal would require repeated HGTs of 11 genes between Thermotogales and Aquificales along all the *Aquifex* genome (Table 1), or a very large HGT of 11 genes, subsequently scattered along the Aquifex genome. Both explanations seem more unlikely. Consequently, we favour the hypothesis of a sin-

Table 1: Position of Aquificales in phylogenies built from single genes present in the concatenated alignment

Position in the genome (locus index)	*Gene name*	*Phylogeny: group neighbouring Aquificales*
8	rpsJ	Thermotogales
11	rplD	*Deinococcus/Thermus*
13	rplB	*Fusobacterium nucleatum*
16	rplV	*Thermoanaerobacter* tengcongensis
17	rpsC	Thermotogales
18	rplP	Planctomycetes
20	rpsQ	*Chloroflexi*
73	rpsK	Planctomycetes
74	rpsM	a clade comprising spirochaetes and *Bacteroidetes/Chlorobi*
123	rpsP	*Bdellovibrio*
226	rpsO	Planctomycetes
287	smb	Thermotogales
461	gatB	Thermotogales
609	hypothetical protein	Clostridiales
712	frr	*Chloroflexi*
735	rpsL2	Thermotogales
792	cycB1	a clade comprising *Thermoanaerobacter tengcongensis* and *Bdellovibrio*
946	rnc	Thermotogales
1478	recR	*Leptospira interogans*
1489	trmD	Thermotogales
1493	dnaG	a clade comprising Spirochaetes and Thermotogales
1645	rpsE	*Deinococcus/Thermus*
1648	rplR	Clostridiales
1649	rplF	Thermotogales
1651	rpsH	a clade comprising Thermotogales and *Deinococcus/Thermus*
1652	rplE	*Actinobacteria*
1654	rplN	*Mycoplasma*
1767	rpsT	*Proteobacteria*
1773	rpmA	*Borrelia*
1777	infC	*Leptospira interogans*
1832	rpsG1	Thermotogales
1878	rpsI	*Desulfotalea psychrophila*
1919	era2	Thermotogales
1933	rplK	Thermotogales
1935	rplA	*Chloroflexi*
1939	rpoB	*Campylobacter jejuni*
1945	rpoC	*Campylobacter jejuni*
2007	rpsB	a clade comprising Thermotogales and *Cyanobacteria*
2032	infB	a clade comprising Proteobacteria, *Bacteroidetes-Chlorobi*, Spirochaetes, Chlamydiales
2042	rplI	a clade comprising delta-Proteobacteria, *Chloroflexi*, and Planctomycetes

Results not unambiguously interpretable are not shown.

gle HGT of the whole rplL-rpoB-rpoC operon from an ancestor of epsilon-Proteobacteria to Aquificales.

Such a hypothesis is relevant to the relative dating of Aquificales and epsilon-Proteobacteria: a transfer from an ancestor of epsilon-Proteobacteria to an ancestor of *Aquifex aeolicus* and *Sulfurihydrogenibium azorense* implies that these ancestors are contemporary. Although in trees of life obtained from rRNAs or concatenated proteins and rooted between Bacteria and Archaea-Eukaryota Aquificales are found very close to the root of Bacteria, the divergence between *Aquifex* and *Sulfurihydrogenibium* should not be more ancient than the divergence of epsilon-Proteobacteria from other Proteobacteria.

A gene-by-gene analysis adds support to the hypothesis that the dominating signal places Aquificales whith Thermotogales. Table 1 shows that, among the 39 gene phylogenies that can be unambiguously interpreted, 11 place Aquificales with Thermotogales while only 2 (RpoB and RpoC) place Aquificales with epsilon-Proteobacteria. The phylogeny of rplL is difficult to interpret, with Aquificales placed close to Delta-proteobacteria and epsilon-Proteobacteria, which might be due to the short length of this

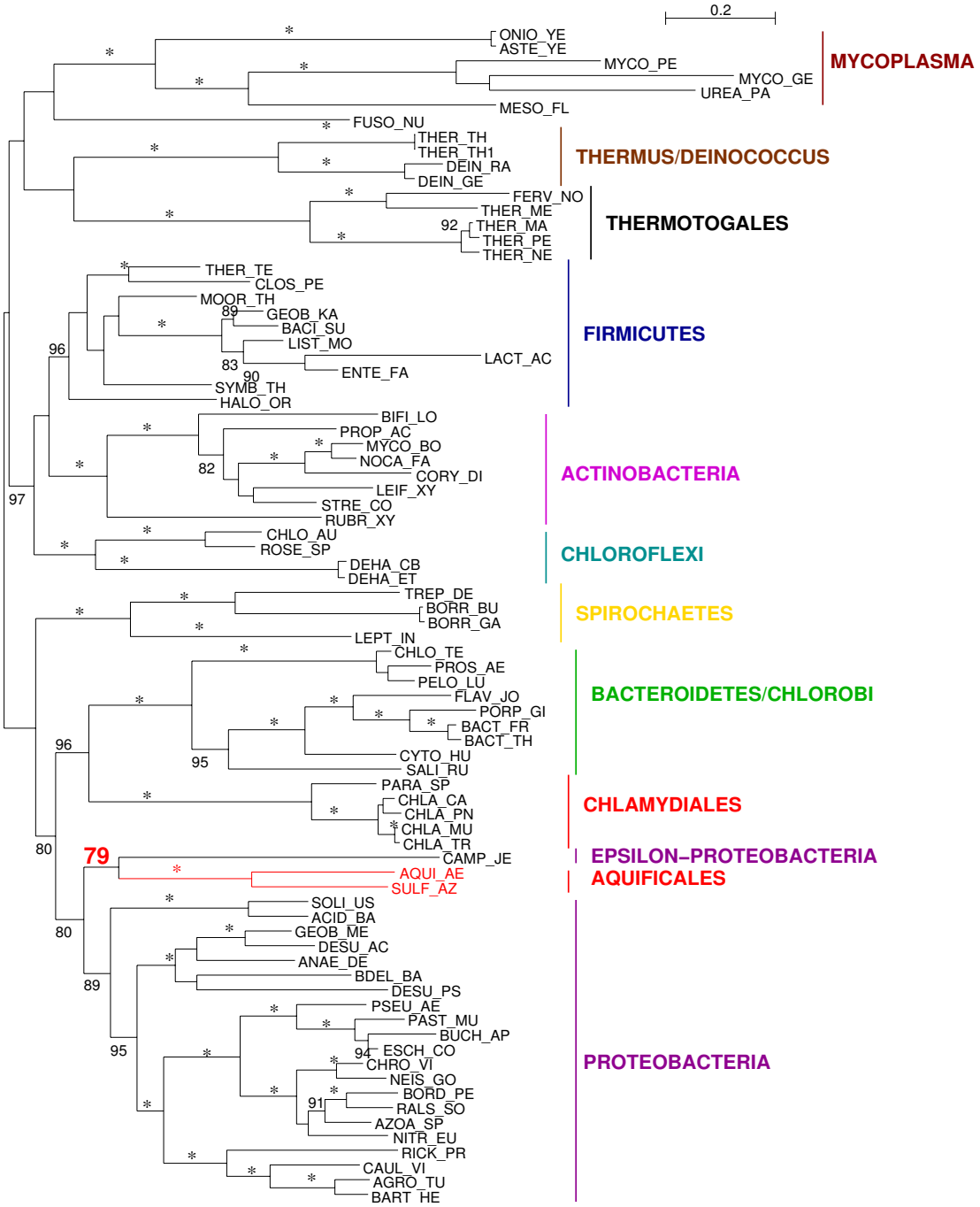


Figure 6
Unrooted tree obtained from the concatenation of rplL-rpoB-rpoC. Colors and symbols as in Fig. 3.

gene (139 sites). Strikingly, 13 genes place Aquificales with Gracilicutes, either close to Planctomycetes, to Spirochaetes, to Bacteroidetes-Chlorobi or to Proteobacteria. A single dominant pattern does not emerge from these gene trees: therefore they do not argue in favour of a specific relationship between Aquificales and a particular group of Gracilicutes. These results rather suggest either uncertainties in phylogenetic reconstruction or repeated horizontal gene transfers between Aquificales and various Gracilicute donors.

In conclusion, the epsilon-proteobacterial signal in the concatenated carefully chosen proteins probably derives from horizontally transferred informational genes, and the Thermotogal signal might be the signal of vertical descent. This conclusion is perfectly congruent with the results from the whole genome analysis. However, the epsilon-Proteobacterial vicinity hypothesis was originally based upon rare genomic changes. How can this hypothesis be reconciled with our conclusions?

The impact of horizontal gene transfers on rare genomic changes

The prevailing cladistic study arguing that Aquificales should be placed as a neighbour to Proteobacteria was performed by Griffiths and Gupta [6], where inserts in 4 genes were found to support this hypothesis. These 4 genes are rpoB, rpoC, alanyl-tRNA synthetase and inorganic pyrophosphatase.

Interestingly, two of these four genes, rpoB and rpoC, are included in our concatenated alignment. Because they are clustered in the *Aquifex aeolicus* genome and display the same non-mainstream phylogenetic signal, we have diagnosed them as resulting from HGT from epsilon-Proteobacteria. Therefore, the two large inserts that Griffiths and Gupta found are no proof of a particular relatedness but rather of a HGT.

The alanyl-tRNA synthetase has not been included in our concatenate because tRNA synthetase genes are known to be extremely prone to HGT [47]. The analysis of the alanyl-tRNA synthetase gene family of the HOGENOM database (family HBG008973), confirms that this gene might not be a good phylogenetic marker. In the tree built from this family with PhyML, *Aquifex aeolicus* is found close to the spirochaete *Leptospira*, together close to Clostridiales, the Planctomycete *Rhodopirellula baltica* is found as a neighbour to Deinococcales (data not shown), among other oddities. All these relations are inconsistent with the tree built from the concatenate and inconsistent with current ideas about bacterial taxonomy. Therefore, using the alanyl-tRNA synthetase gene family to resolve bacterial phylogeny appears inadequate.

Finally, the inorganic pyrophosphatase tree as retrieved from HOGENOM (family HBG000457) shows *Aquifex aeolicus* inside Proteobacteria, close to Alpha-proteobacteria, which are not monophyletic. It appears that this gene family has undergone a duplication (Cyanobacteria are represented twice in the tree in widely separated positions) as well as horizontal gene transfers (Archaea are clustered in two groups widely separated in the tree, as well as Chlamydiales). Overall, the history of inorganic pyrophosphatase is probably too complex to be used as a marker of species relationships.

Consequently, the rare genomic changes that were used to argue for a specific relatedness between Aquificales and Proteobacteria most likely come from HGT between these two clades, as already observed in the above analyses (Fig. 1 for instance).

The fact that the outer membrane of *Aquifex* closely resembles the outer membrane of other Proteobacteria was also used [4] to argue that Aquificales are more closely related to Proteobacteria than to Thermotogales. It is unclear why this character would be particularly immune to HGT; the outer membrane most likely possesses a strong adaptive value, so that the transfer of the operational genes coding for such a structure could be positively selected and rise to fixation in a species. Given the very high rate of HGT seen in *Aquifex* genome, it is not unreasonable to assume that the proteobacterial type of outer membrane might have been transferred to Aquificales. Similarly, the close relationship found between epsilon-Proteobacteria and Aquificales in trees based on cytochromes b and c might also come from a HGT of a whole operon, as concluded by Schutz *et al*. [48]. On the contrary, our counting analysis confirms that informational genes are less prone to HGT than operational genes, and their signal clusters Aquificales and Thermotogales.

Further difficulties to resolve the tree of Bacteria

A possible approach to uncover a putative species tree of Bacteria, or at least a tree for a core set of bacterial genes, would be to remove transferred genes from a dataset, concatenate all genes that have not been detected as having been transferred, and use them to build a phylogenetic tree. Such an approach would be expected to yield better trees, with higher bootstrap supports. However, the phylogeny obtained on the concatenate in the same conditions as before (without recoding) but after removal of the rplL-rpoB-rpoC genes does not show a significantly better support for most of its nodes than the phylogeny shown in Fig. 3 (average bootstrap support for the tree without the three genes, 90.9, and for the tree with all genes, 90.3; p-value = 0.17 with a Student paired t-test, p-value = 0.288 with a Wilcoxon signed rank test). This is probably due to the fact that bootstrap supports increase with the number

of characters; the length parameter therefore counters the expected positive effect associated with the removal of discordant signal. Topologically, both trees are highly congruent, with the main noticeable difference being the placement of *Fusobacterium nucleatum*, which leaves its position as sister-group to Thermotogales and Aquificales in Fig. 3 to nest inside the Firmicutes as a sister group to *Mycoplasma*. This placement might stem from a long branch attraction, as both *Mycoplasma* and *Fusobacterium* have long terminal branches, or alternatively might reveal the true history of *Fusobacterium nucleatum*, as suggested by Mira and co-workers [49]. Certainly this organism deserves further study, possibly with techniques such as those that were used in this article.

It is interesting to note that the removal of genes thought to have been transferred has not improved the phylogeny. A most promising avenue for further research in deep phylogenies would probably involve the development of models explicitly taking into account HGT, as proposed by Suchard [50] or, in other contexts, by Edwards, Liu and Pearl [51,52] and Ané *et al*. [53]. HGTs should be modelled as a genuine biological phenomenon on equal footing with vertical descent to represent the evolution of bacterial genomes. The resulting species tree would correspond to the history of those genome parts that have been vertically inherited at any time during evolution. The vertically inherited portions of a genome at a given time need not be vertically inherited at all time, so that a species tree could be inferred as long as, at any time, some vertical signal could be recovered.

Another additional difficulty might be that the gene is not necessarily the atomic unit of transfer: transfers may affect only parts of a gene, through recombination. In this respect, the analysis of Figure 5 reveals a striking pattern in the Initiation Factor 2 gene (infB, green line). In this large gene (the *Aquifex aeolicus* protein is 805 amino-acids long), the curve of the difference in log-likelihoods between the epsilon-proteobacterial and the thermotogal positions of Aquificales first decreases for about half its length, and then increases. This pattern is suggestive of a recombination event inside the gene.

To test for recombination, we divided the infB alignment in two at the point where this curve changes trend and built phylogenetic trees for both partial alignments (Fig. 7). In the first resulting tree, Aquificales plus *Fusobacterium nucleatum* make together a sister group to Thermotogales plus *Deinococcus/Thermus*. In the second tree, Aquificales are a sister group to a subclade of Firmicutes. These two branchings are consistent with the slope of the curve of Fig. 5, first descending, as Aquificales are close to Thermotogales, and then ascending, as Aquificales are far from Thermotogales. To assess whether the differences in the topologies were significant, Consel was used [42] on these last two trees. The first part of the alignment strongly rejected the tree obtained for the second part (AU test p-value: 4.10^{-36}; SH and KH p-value: 0), and *vice versa* (AU test p-value: 1.10^{-06}; SH and KH p-value: 0). Therefore a strong signal for recombination within the gene infB is found, possibly between Firmicutes and Aquificales.

This indicates that the unit of transfer between Bacteria is not necessarily the gene, but can also be parts of a gene. Models aiming at resolving the bacterial tree may need to take this additional complexity into account.

Conclusion

Overall, the signal in favour of a close relationship between Aquificales and epsilon-Proteobacteria has been shown to be coming from a lateral transfer and not vertically inherited, both in protein phylogenies and in cladistic analyses. A large HGT involving three consecutive genes encoding two RNA polymerase subunits and a ribosomal protein has been detected. This large gene transfer between epsilon-Proteobacteria and Aquificales can be understood in terms of a shared ecological niche: some epsilon-Proteobacteria are indeed found in hyperthermophilic environments [54].

The present single-gene analyses suggested that gene transfers may have frequently occurred between Aquificales and various Gracilicutes and Proteobacteria in particular, which explains why cladistic analyses of rare genomic changes or of domain contents often place *Aquifex* inside Gracilicutes.

Bacterial phylogeny is crucial to understand the evolution of the biosphere, as it provides a backbone permitting to integrate the evolution of life as revealed from molecular phylogenies with the history of the earth, as dug up by geology. There is no doubt that HGT has played a major role in the evolution of Prokaryotes, to the point that there might be no gene that has never undergone HGT; however a few gene families may have seldom been transferred, and they might bear sufficient signal to unveil the vertical history of the genome, provided powerful computational methods modelling both gene transfers and intragenic recombination are developed.

Nonetheless, because Aquificales are often found grouped with Thermotogales, and because this phylogenetic signal does not seem to result from known artefacts such as long branch attraction or compositional bias, if there is a species tree in Bacteria, Aquificales are to be considered as a sister group to Thermotogales. This clarification does not dramatically affect the scenario for the evolution of life proposed by Cavalier-Smith [9], except that Aquificales diverged earlier than proposed. However the present

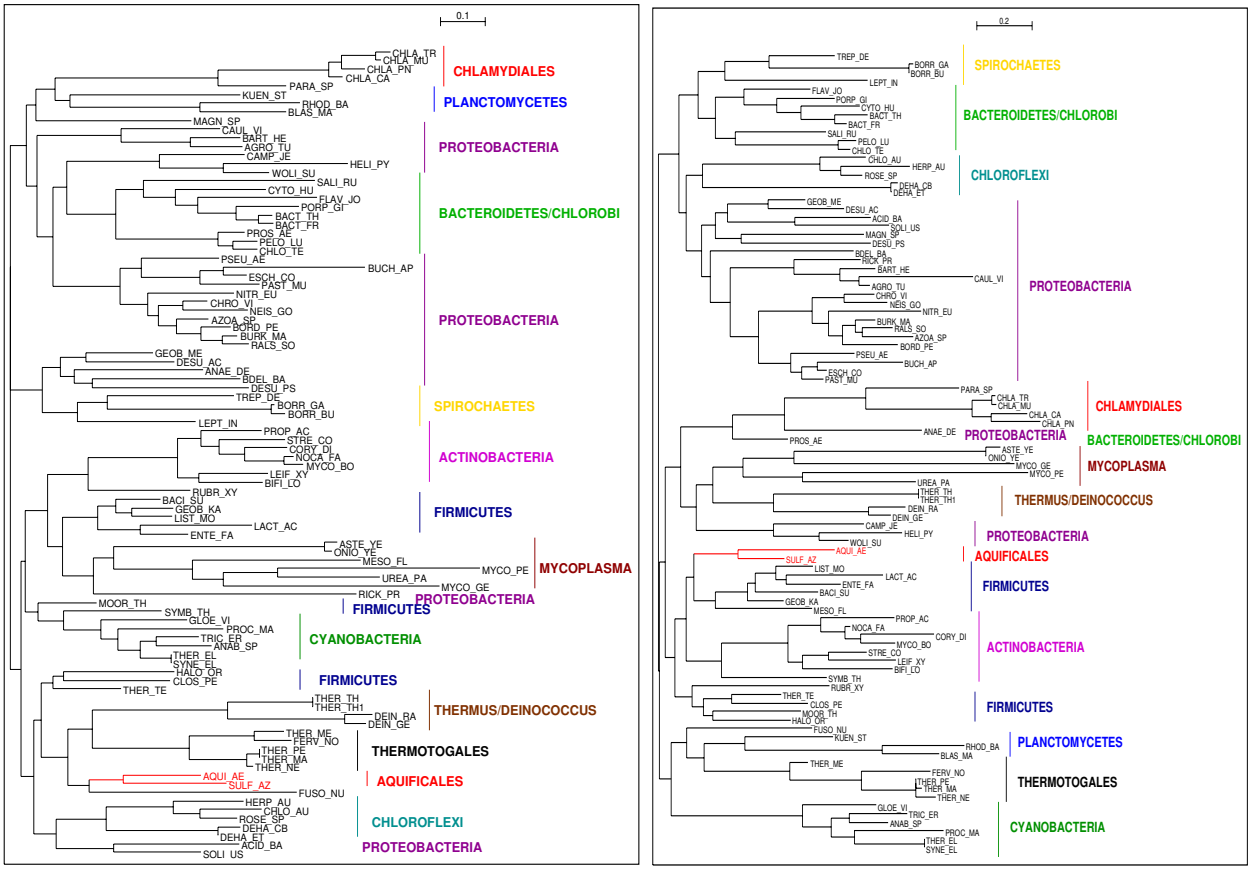


Figure 7
Unrooted trees corresponding to the infB gene. Left: tree corresponding to the first 301 sites. Right: tree corresponding to the remaining 246 sites. Colors as in Fig. 3.

results question the methodology used to build this scenario because the rare genomic changes method requires that HGT does not affect used marker genes. In the case of the Aquificales, we have shown that this requirement is not fulfilled.

Methods

Whole phylome analysis

In order to get a whole genome view of Aquificales phylogenetic relationships, we queried the HOGENOM database (release 03, October 2005) using the TreePattern program in FamFetch [55]. HOGENOM is a database that clusters sequences from whole genomes into homologous gene families, and builds trees based on these families with PhyML using a gamma law with 4 classes of substitution rates, with estimated alpha parameter and proportion of invariable sites. Trees corresponding to all 892 families in which there was a sequence from *Aquifex aeolicus* were automatically analysed, and each sequence from *Aquifex* was classified according to what group of species appeared as its closest neighbour, not taking into account branch support or branch length. This gave counts of *Aquifex* genes found close to Thermotogales, Firmicutes, epsilon-Proteobacteria, *etc*... Cases where *Aquifex* genes were found close to a non-monophyletic group of species were discarded, which left 578 gene trees. These counts were further classified into two functional categories, "informational genes" and "non-informational genes", through TIGRFAM annotations [56]. A functional category could be determined for 351 families. "Informational genes" were genes classified in TIGRFAMs whose function was part of "Transcription", "DNA metabolism", "Protein synthesis"; "non-informational genes" were those whose role was part of other major functional classes.

Concatenate assembly

Nearly universal gene families which had only one copy per genome were used to minimize problems of ill-

defined orthology. Consequently, gene families from the HOGENOM database of families of homologous genes (release 03, October 2005) that displayed a wide species coverage with no or very low redundancy in all species were selected. This provided 70 gene families. Sequences from representative genomes from Archaea were retrieved from these families, and sequences from genomes not present in the release 03 of HOGENOM but whose phylogenetic position was interesting were included in the families. These studied genomes are listed in Additional files 1, 2 and 3 and were downloaded from the Joint Genome Institute [57], The Institute for Genomic Research [58] or the National Center for Biotechnology Information [59], and were searched for homologous genes using BLAST [60]; only the best hit was retrieved. The gene families were subsequently aligned using MUSCLE v3.52 [61] and submitted to a phylogenetic analysis using the NJ algorithm [62] with Poisson distances as implemented in Phylo_Win [63]. During this step, families in which there seemed to be a gene transfer between a bacterial species and Archaea were discarded, as well as amino-acid synthetases, which are known to be prone to HGT [47]. In the rare families where there were two sequences from the same species, the sequence showing the largest terminal branch length or whose position was most at odds with the NCBI classification was discarded. This whole process provided 56 gene families and 94 bacterial species. Only bacterial sequences were used in the rest of the study, because our focus is on the bacterial phylogeny itself. The 56 families were submitted to Gblocks [64] to discard parts of the alignments that were unreliable, but using a non-stringent site selection, because the subsequent analyses should permit to sort biased from genuine signal. Consequently, the following Gblocks parameters were used: the minimum numbers of sequences used to define a conserved or a flanking position were set at 50% of the total number of sequences, the minimum length of a block was set at 2 sites, and all positions could be kept by the algorithm, even if they contained gaps. The resulting alignments were then concatenated using ScaFos [65], following the order of genes along the *Aquifex aeolicus* genome. The amount of missing data was low, reaching 21% at its maximum in *Thermotoga petrophila*.

Phylogenetic analyses

A phylogenetic tree was built from the concatenate under the Maximum Likelihood criterion using PhyML v.2.4.4 [66] with the JTT model [67], and a discretized gamma law with 4 categories to model evolutionary rate variation. This first tree was used to compute site-specific evolutionary rates using BppML from the Bio++ package [68], which allowed for the removal of saturated sites. A new tree was built using this refined alignment, with the same parameters plus an estimated proportion of invariant sites and with a non-parametric bootstrap analysis (200 replicates), and was used as a reference for the rest of the work. An estimated proportion of invariant sites was not used in the previous analysis because it had not been implemented in the used version of Bio++. Noticeably, the topology was found to be unchanged when Aquificales were removed from the alignment and the tree re-computed. Similarly, the topology was nearly identical when two free-living espsilon-Proteobacteria (*Sulfurovum* NBC37-1 and thermophilic *Nitratiruptor* SB155-2 [40],) were added, and the tree recomputed with PhyML v3.0; for this tree, the minimum of SH-like and chi2-based support was computed instead of bootstrap support [69]. An additional test was performed to assess the impact of compositional heterogeneity as well as saturation: the alignment without saturated sites was recoded in 4 categories [70,37]. In this recoding, aromatic (FWY) and hydrophobic (MILV) amino-acids were grouped in a single state, basic amino-acids (HKR) in another, acidic (DENQ) amino acids in one more state, and the fourth state contained all other amino acids (AGPST) to the exception of cysteine which was coded as missing data. The recoded alignment was subjected to a phylogenetic analysis with the GTR model [71], an estimated proportion of invariant sites, a gamma law discretized in 8 categories with its alpha parameter estimated, and 200 bootstrap replicates.

The tree without the Aquificales was used as a scaffold upon which all possible Aquificales positions were tried in turn. The likelihoods for each of these positions were computed using BppML from the Bio++ package. Evolutionary rates per site as well as likelihoods per site were simultaneously inferred. Site evolutionary rates were obtained by computing the average of the gamma law rate categories weighted by their posterior probabilities.

The tree containing only the rplL-rpoB-rpoC genes was obtained with PhyML as described above and with a non-parametric bootstrap analysis based upon 500 replicates.

Individual gene trees were built using PhyML with the same parameters as above except that the gamma law was discretized in 8 categories.

Concatenate segmentation and HGT identification

We wanted to know which was the most likely segmentation in two segments of the alignment according to site likelihoods for all topologies. It was computed using Sarment [72] with the Maximum Predictive Partitioning algorithm [43]. This algorithm was input a matrix containing the site log-likelihoods for all 181 topologies tested (obtained by placing the Aquificales in all possible positions in the backbone bacterial phylogeny) and for the whole alignment. The best log-likelihood of a given segmentation is the sum of the best log-likelihoods of its

segments, that are computed as follows: on a segment, for each of the 181 topologies tested, the log-likelihood of a topology is the sum of all site log-likelihoods on the alignment. This procedure produces 181 log-likelihoods, the maximum of which is the best log-likelihood of this segment. Once this maximum is found, it clearly associates a most likely topology to each segment of the alignment. All statistical analyses were done with the seqinR package [73] in R [74].

Abbreviations

HGT: Horizontal Gene Transfer; rRNA: ribosomal Ribo-Nucleic Acid; indel: insertion-deletion; MPP: Maximum Predictive Partitioning.

Authors' contributions

MG and BB designed the study. LG performed the segmentation analysis, and BB performed the other experiments. BB wrote most of the manuscript, which was improved by LG and MG.

Additional material

Additional file 1

The list of species used in the study, and their abbreviated names as found in the figures of the article.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-272-S1.xls]

Additional file 2

Unrooted phylogenetic tree of Bacteria obtained after the addition of two free-living epsilon-Proteobacteria,* Sulfurovum *NBC37-1 and thermophilic* Nitratiruptor *SB155-2.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-272-S2.jpeg]

Additional file 3

The list of 56 HOGENOM gene families used to estimate species trees, with the corresponding function description.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-272-S3.xls]

Acknowledgements

We wish to thank Vincent Daubin, Anamaria Necşulea, Leonor Palmeira and Sophie Abby for valuable discussions and help with R and the data. Preliminary sequence data was obtained from The Institute for Genomic Research through the website at http://www.tigr.org. Sequencing of *Sulfurihydrogenibium azorense* Az-Fu was accomplished with support from NSF. Sequencing of *Thermotoga neapolitana* DSM 4359 was accomplished with support from DOE. This work was supported by Action Concertée Incitative IMPBIO. We thank the Centre de Calcul de l'IN2P3 for providing computer resources. Bastien Boussau acknowledges a PhD scholarship from the Centre National de la Recherche Scientifique.

References

1. Wolf YI, Rogozin IB, Grishin NV, Koonin EV: **Genome trees and the tree of life.** *Trends Genet* 2002, **18:**472-479.
2. Deeds EJ, Hennessey H, Shakhnovich EI: **Prokaryotic phylogenies inferred from protein structural domains.** *Gen Res* 2005, **15:**393-402.
3. Klenk HP, Meier TD, Durovic P, Schwass V, Lottspeich F, Dennis PP, Zillig W: **RNA polymerase of Aquifex pyrophilus: implications for the evolution of the bacterial rpoBC operon and extremely thermophilic bacteria.** *J Mol Evol* 1999, **48:**528-541.
4. Cavalier-Smith T: **The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification.** *Int J Syst Evol Microbiol* 2002, **52:**7-76.
5. Coenye T, Vandamme P: **A genomic perspective on the relationship between the Aquificales and the epsilon-Proteobacteria.** *Syst Appl Microbiol* 2004, **27:**313-322.
6. Griffiths E, Gupta RS: **Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales.** *Int Microbiol* 2004, **7:**41-52.
7. Cavalier-Smith T: **Rooting the tree of life by transition analyses.** *Biol Direct* 2006, **1:**19-19.
8. Kunisawa T: **Dichotomy of major bacterial phyla inferred from gene arrangement comparisons.** *J of Theor Biol* 2006, **239:**367-375.
9. Cavalier-Smith T: **Cell evolution and Earth history: stasis and revolution.** *Philos T R Soc B* 2006, **361:**969-1006.
10. Jain R, Rivera MC, Lake JA: **Horizontal gene transfer among genomes: the complexity hypothesis.** *P Natl A Sci USA* 1999, **96:**3801-3806.
11. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading.** *Syst Zool* 1978, **27:**401-410.
12. Brinkmann H, Giezen M van der, Zhou Y, Poncelin De Raucourt G, Philippe H: **An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics.** *Syst Biol* 2005, **54:**743-757.
13. Weisburg WG, Giovannoni SJ, Woese CR: **The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction.** *Syst Appl Microbiol* 1989, **11:**128-134.
14. Foster PG, Hickey DA: **Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions.** *J Mol Evol* 1999, **48:**284-290.
15. Schopf JW: **Fossil evidence of Archaean life.** *Philos T R Soc B* 2006, **361:**869-85.
16. Galtier N, Lobry JR: **Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes.** *J Mol Evol* 1997, **44:**632-636.
17. Zeldovich KB, Berezovsky IN, Shakhnovich EI: **Protein and DNA sequence determinants of thermophilic adaptation.** *PLoS Comput Biol* 2007, **3:**e5-e5.
18. Rokas A, Holland PW: **Rare genomic changes as a tool for phylogenetics.** *Trends Ecol Evol* 2000, **15:**454-459.
19. Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Swanson RV: **The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*.** *Nature* 1998, **392:**353-358.
20. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM: **Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima.** *Nature* 1999, **399:**323-329.
21. **The HOGENOM database** [http://pbil.univ-lyon1.fr/databases/hogenom3.html]
22. Zar JH: *Biostatistical Analysis* 4th edition. Upper Saddle River: Prentice Hall; 1999.
23. Leigh JW, Susko E, Baumgartner M, Roger AJ: **Testing congruence in phylogenomic analysis.** *Syst Biol* 2008, **57:**104-115.
24. Bapteste E, Susko E, Leigh J, Ruiz-Trillo I, Bucknam J, Doolittle WF: **Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny.** *Mol Biol Evol* 2008, **25:**83-91.

25. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol* 1994, **39:**306-314.
26. Lartillot N, Brinkmann H, Philippe H: **Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model.** *BMC Evol Biol* 2007, **1:**S4-S4.
27. Brochier C, Philippe H: **Phylogeny: a non-hyperthermophilic ancestor for bacteria.** *Nature* 2002, **417:**244-244.
28. Battistuzzi FU, Feijao A, Hedges SB: **A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land.** *BMC Evol Biol* 2004, **4:**44-44.
29. Bern M, Goldberg D: **Automatic selection of representative proteins for bacterial phylogeny.** *BMC Evol Biol* 2005, **5:**34-34.
30. Daubin V, Gouy M, Perrière G: **A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history.** *Genome Res* 2002, **12(7):**1080-1090.
31. Wagner M, Horn M: **The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance.** *Curr Opin Biotech* 2006, **17:**241-249.
32. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311:**1283-1287.
33. Beiko RG, Harlow TJ, Ragan MA: **Highways of gene sharing in prokaryotes.** *P Natl A Sci USA* 2005, **102:**14332-14337.
34. Ababneh F, Jermiin LS, Ma C, Robinson J: **Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences.** *Bioinformatics* 2006, **22:**1225-31.
35. Bowker AH: **A test for symmetry in contingency tables.** *J Am Stat Assoc* 1948, **43:**572-574.
36. Kreil DP, Ouzounis CA: **Identification of thermophilic species by the amino acid compositions deduced from their genomes.** *Nucleic Acids Res* 2001, **29:**1608-15.
37. Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H: **Detecting and overcoming systematic errors in genome-scale phylogenies.** *Syst Biol* 2007, **56:**389-399.
38. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE: **Whole-genome analysis of photosynthetic prokaryotes.** *Science* 2002, **298:**1616-1620.
39. Susko E, Roger AJ: **On reduced amino Acid alphabets for phylogenetic inference.** *Mol Biol Evol* 2007, **24:**2139-50.
40. Nakagawa S, Takaki Y, Shimamura S, Reysenbach AL, Takai K, Horikoshi K: **Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens.** *Proc Natl A Sci USA* 2007, **29:**12146-12150.
41. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Syst Biol* 2002, **51:**492-508.
42. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17:**1246-1247.
43. Guéguen L: **Segmentation by maximal predictive partitioning according to composition biases.** *Computational Biology, LNCS, 2066* 2001:32-45.
44. Newman AJ, Linn TG, Hayward RS: **Evidence for co-transcription of the RNA polymerase genes rpoBC with a ribosomal protein gene of escherichia coli.** *Mol Gen Genet* 1979, **169:**195-204.
45. Yamamoto M, Nomura M: **Contranscription of genes for RNA polymerase subunits beta and beta' with genes for ribosomal proteins in Escherichia coli.** *P Natl A Sci USA* 1978, **75:**3891-3895.
46. Iyer LM, Koonin EV, Aravind L: **Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer.** *Gene* 2004, **23:**73-88.
47. Wolf YI, Aravind L, Grishin NV, Koonin EV: **Evolution of aminoacyl-tRNA synthetases-analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events.** *Gen Res* 1999, **9:**689-710.
48. Schütz M, Brugna M, Lebrun E, Baymann F, Huber R, Stetter KO, Hauska G, Toci R, Lemesle-Meunier D, Tron P, Schmidt C, Nitschke W: **Early evolution of cytochrome bc complexes.** *J Mol Biol* 2000, **300:**663-675.
49. Mira A, Pushker R, Legault BA, Moreira D, Rodríguez-Valera F: **Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics.** *BMC Evol Biol* 2004, **4:**50-50.
50. Suchard MA: **Stochastic models for horizontal gene transfer: taking a random walk through tree space.** *Genetics* 2005, **170:**419-31.
51. Edwards SV, Liu L, Pearl DK: **High-resolution species trees without concatenation.** *P Natl A Sci USA* 2007, **104:**5936-5941.
52. Liu L, Pearl DK: **Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions.** *Syst Biol* 2007, **56:**504-14.
53. Ané C, Larget B, Baum DA, Smith SD, Rokas A: **Bayesian estimation of concordance among gene trees.** *Mol Biol Evol* 2007, **24:**412-26.
54. Nakagawa S, Takaki Y, Shimamura S, Reysenbach AL, Takai K, Horikoshi K: **Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens.** *P Natl A Sci USA* 2007, **104:**12146-12150.
55. Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G: **Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases.** *Bioinformatics* 2005, **21:**2596-2603.
56. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic Acids Res* 2003, **31:**371-373.
57. **The Joint Genome Institute** [http://www.jgi.doe.gov/]
58. **The Institute for Genomic Research** [http://www.tigr.org/]
59. **The National Center for Biotechnology Information** [http://www.ncbi.nlm.nih.gov/]
60. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
61. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792-1797.
62. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
63. Galtier N, Gouy M, Gautier C: **SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny.** *Comput Appl Biosci* 1996, **12:**543-548.
64. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17:**540-552.
65. Roure B, Rodriguez-Ezpeleta N, Philippe H: **SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics.** *BMC Evol Biol* 2007, **7(Suppl 1):**S2-S2.
66. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696-704.
67. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appli Biosci* 1992, **8:**275-282.
68. Dutheil J, Gaillard S, Bazin E, Glemin S, Ranwez V, Galtier N, Belkhir K: **Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics.** *BMC Bioinformatics* 2006, **7:**188-188.
69. Anisimova M, Gascuel O: **Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative.** *Syst Biol* 2006, **55:**539-552.
70. Hrdy I, Hirt RP, Dolezal P, Bardonová L, Foster PG, Tachezy J, Embley TM: **Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I.** *Nature* 2004, **432:**618-622.
71. Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20:**86-93.
72. Guéguen L: **Sarment: Python modules for HMM analysis and partitioning of sequences.** *Bioinformatics* 2005, **21:**3427-3428.
73. Charif D, Thioulouse J, Lobry JR, Perrière G: **Online synonymous codon usage analyses with the ade4 and seqinR packages.** *Bioinformatics* 2005, **21:**545-547.
74. R Development Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing Vienna, Austria; 2005. ISBN 3-900051-07-0

4

Improving Methods of Phylogenetic Reconstruction

The preceding article showed that compositional heterogeneity was a very difficult problem for phylogenetic reconstruction, as it cannot be erased easily. The best approach to reconstructing phylogeny when there are compositional biases involves better models of evolution.

These better models of evolution are non-homogeneous (see section 2.7.2), and were thought to be uneasy to work with, because they render the whole process of evolution irreversible. Here we show that this irreversibility does not prevent from using classical and efficient algorithms. As a proof-of-concept, I programmed nhPhyML, an algorithm that implements a model designed to deal with compositional heterogeneity.

This article has been published in *Systematic Biology*.

Syst. Biol. 55(5):756–768, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600975218

Efficient Likelihood Computations with Nonreversible Models of Evolution

BASTIEN BOUSSAU AND MANOLO GOUY

Laboratoire de Biométrie et Biologie Evolutive (UMR 5558); CNRS; Université Lyon 1, 43 boulevard 11 nov 1918, 69622, Villeurbanne Cedex, France; E-mail: boussau@biomserv.univ-lyon1.fr (B.B.)

Abstract.—Recent advances in heuristics have made maximum likelihood phylogenetic tree estimation tractable for hundreds of sequences. Noticeably, these algorithms are currently limited to reversible models of evolution, in which Felsenstein's pulley principle applies. In this paper we show that by reorganizing the way likelihood is computed, one can efficiently compute the likelihood of a tree from any of its nodes with a nonreversible model of DNA sequence evolution, and hence benefit from cutting-edge heuristics. This computational trick can be used with reversible models of evolution without any extra cost. We then introduce nhPhyML, the adaptation of the nonhomogeneous nonstationary model of Galtier and Gouy (1998; Mol. Biol. Evol. 15:871–879) to the structure of PhyML, as well as an approximation of the model in which the set of equilibrium frequencies is limited. This new version shows good results both in terms of exploration of the space of tree topologies and ancestral G+C content estimation. We eventually apply it to rRNA sequences slowly evolving sites and conclude that the model and a wider taxonomic sampling still do not plead for a hyperthermophilic last universal common ancestor. [Efficient algorithm; LUCA; maximum likelihood; molecular phylogeny; nonreversible model of evolution; PhyML; origin of life; root of life.]

Research in molecular phylogeny aims at reconstructing historical relations between genes or species while trying to capture the true nature of the evolutionary process itself. Both can be estimated at the same time through the use of statistical modeling. Maximum likelihood or the Bayesian framework permit the estimation of parameters of the evolutionary model such as the transition/transversion ratio, the equilibrium base composition, and the tree itself, topology and branch lengths included. Optimizing all these parameters is computationally intensive: the number of possible topologies increases factorially with the number of taxa considered, which makes it necessary to use heuristics when exploring the space of tree topologies. Most recent algorithms (e.g., PhyML [Guindon and Gascuel, 2003], RAxML [Stamatakis et al., 2005]) are able to find trees with excellent likelihood scores for hundreds of sequences, but only with reversible models of evolution. All these reversible models are homogeneous and stationary, i.e., suppose that state evolution is constant all over the tree. If this hypothesis were true, sequences sharing a common ancestor would have the same expected base frequencies.

More precisely, a process of evolution is homogeneous when the state distribution probability simply depends on the time separating it from a given past state distribution probability and not on the branch in the tree: homogeneity is the feature of a process of evolution that is constant in pattern over the whole tree. On the other hand, stationarity is the feature of a process of evolution that keeps the state distribution probability constant over the whole tree: the probability to draw a given state is the same wherever on the tree the sampling is done. A process can be stationary and not be homogeneous, as is the case for Ziheng Yang's codon model in which the nonsynonymous-to-synonymous ratio varies across branches while codon equilibrium frequencies remain constant all over the tree (Yang, 1998). On the contrary, nonstationarity induces nonhomogeneity, as the process of evolution depends upon the equilibrium frequencies.

The analysis of extant sequences shows that homologous genes vary widely in their composition. As they all stem from a common ancestor, this evidences that sequence evolution is at least not stationary: two sequences in two different species or at two different periods evolve towards different compositions.

The use of nonhomogeneous and nonstationary models that account for this variability in evolution permits minimizing compositional biases and hence improving phylogenetic reconstructions (Galtier and Gouy, 1998; Tarrio et al., 2001; Herbeck et al., 2005). Unfortunately, removing the homogeneity and stationarity hypotheses implies abandoning reversibility, and hence prevents one from using the most efficient algorithms, when those particularly variable-rich models would most eagerly need it.

In this article we show in the general case that it is possible to use recent algorithms with nonreversible models of sequence evolution. We first explain how the likelihood of a tree is computed and how the reversibility property is used in recent heuristics to avoid dispensable calculations during tree space search. We then prove that the same computational trick can be used with nonreversible models of evolution through a reorganization of the way likelihood is computed. As reversible models of evolution can also be used in this framework, this work can be considered as a generalization of the usual formulas, which considerably broadens the amount of models that can be used for a phylogenetic analysis, as all nonhomogeneous as well as nonstationary models can be used.

Eventually we report nhPhyML, the first adaptation of PhyML (Guindon and Gascuel, 2003), a very fast and efficient algorithm, to a nonreversible model of evolution; namely, the implementation of Tamura's model at each branch of a tree introduced by Galtier and Gouy (Tamura, 1992; Galtier and Gouy, 1998). We also introduce nhPhyML-Discrete, an approximation of nhPhyML. This implementation shows better performance than nhPhyML in the exploration of the space of tree

topologies and similar accuracy in the estimation of the ancestral G+C content. Eventually, we apply nhPhyML to ribosomal RNA slowly evolving sites and conclude that a wider taxonomic sampling than in Galtier et al. (1999) still does not support a hyperthermophilic last universal common ancestor.

Computing the Likelihood of a Tree from Any Node under a Nonreversible Model of DNA Sequence Evolution

Computing the Likelihood of a Tree

We first explain how one computes the likelihood of a phylogenetic tree with DNA sequences using the following example (Fig. 1).

Most commonly, sites are supposed to evolve independently of each other: a site does not depend on its neighbors' states but only on its past state. As a consequence, the likelihood of a tree for a whole sequence is obtained by multiplying all the likelihoods obtained at single sites.

The likelihood L_s of the tree given in Figure 1 for a single site s is computed as follows:

$$L_s = \sum_{x\in\Omega}\Bigg(P(R=x)\times\sum_{z\in\Omega}[P_{xz}(l_A, v_A)L_{s,low(RA)}(A=z)] \times\sum_{y\in\Omega}\Bigg\{P_{xy}(l_U, v_U)\sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)}(B=q)] \times\sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)]\Bigg\}\Bigg) \quad (1)$$

where $P_{xy}(l_A, v_A)$ is the probability for base x to change into base y along a branch of length l_A and other evolutionary parameters v_A, $P(R=x)$ is the probability to have base x at the root R, and $\Omega=\{A, T, C, G\}$ is the set of possible DNA bases. $L_{s,low(RA)}(A=z)$ is the lower conditional likelihood (Felsenstein, 1981) of observing the data downstream from branch RA conditionally on the underlying subtree and on having base z at node A. For each subtree, one can define four conditional likelihoods, one for each DNA base. Once these conditional likelihoods have been computed for a subtree, as long as its topology and branch lengths do not change, they can be re-used if one moves the whole subtree around the topology. This property is used in recent heuristics to search for the most likely phylogenetic tree. These conditional likelihoods are defined as lower, in the sense that they do not contain the root.

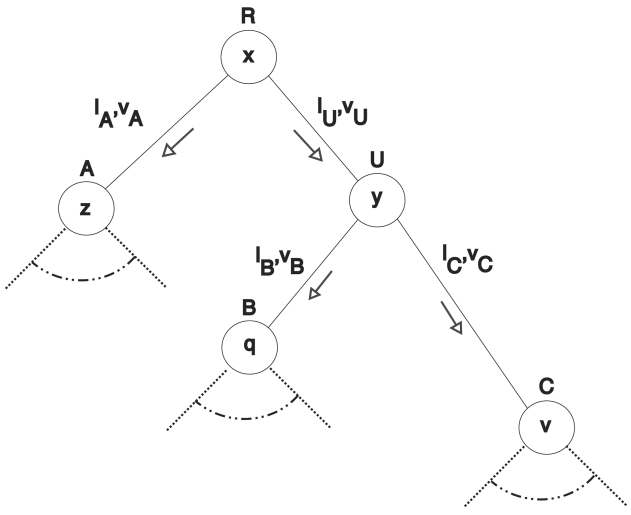


FIGURE 1. Example rooted tree for likelihood computation. This tree is composed of a root R, an internal node U, three other nodes or leaves A, B, and C, and four branches of length l_A, l_B, l_C, l_U, and other evolutionary parameters v_A, v_B, v_C, v_U. We are here interested in the likelihood of the tree for a single site. The internal node states are unknown and then represented as variables x at node R, y at node U, z at node A, q at node B, and v at node C. Arrows represent the evolutionary direction, from the root of the tree to its leaves.

Lower conditional likelihoods are defined recursively. For a leaf C:

$$L_{s,low(UC)}(C=v)=\begin{cases}1 \text{ if base v is at site s of leaf C}\\ 0 \qquad\qquad \text{otherwise}\end{cases} \quad (2)$$

And for a subtree whose root is in U:

$$L_{s,low(RU)}(U=y)=\sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)}(B=q)] \times\sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)] \quad (3)$$

Computing the Likelihood When the Model of Evolution Is Reversible

Reversibility.—When computing the likelihood of a phylogenetic tree, a root R must be specified. If the model is homogeneous and reversible, the process of evolution is stationary: wherever the root is, its base proportions are the same, i.e., they are the equilibrium frequencies of the process, noted π: $P(R=x)=\pi_x$. (1) can be rewritten:

$$L_s=\sum_{x\in\Omega}\Bigg(\pi_x\sum_{y\in\Omega}\Bigg\{P_{xy}(l_U, v_U) \times\sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)}(B=q)] \times\sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)]\Bigg\} \times\sum_{z\in\Omega}[P_{xz}(l_A, v_A)L_{s,low(RA)}(A=z)]\Bigg) \quad (4)$$

Reversibility means that, averaged over the whole sequence, the flux from one base to another is equal to the flux from this other base back to the first one:

$$\pi_x P_{xy}(l, v)=\pi_y P_{yx}(l, v)$$

Supposing the model used is reversible, as is the case with most current models of DNA sequence evolution, we can rewrite the likelihood in (4) as:

$$L_s = \sum_{y\in\Omega}\Bigg(\pi_y \sum_{x\in\Omega}\Bigg\{P_{yx}(l_U, v_U) \times \sum_{z\in\Omega}[P_{xz}(l_A, v_A)L_{s,low(RA)}(A=z)]\Bigg\} \times \sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)}(B=q)] \times \sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)]\Bigg) \quad (5)$$

Expression (5) can be read as if the root was placed at node U. The root can therefore be placed at any node, on any branch of the tree, a property named "pulley principle" by Felsenstein (1981), and widely used in heuristics to find most likely trees. Considering Figure 1, this makes arrows meaningless.

The possibility to place the root of the tree wherever is needed is thoroughly used in recent heuristics to the problem of the most likely phylogenetic tree. Those heuristics usually explore the space of tree topologies by applying local rearrangements: "nearest neighbor interchange" (NNI) swaps two subtrees around an internal edge (used in PhyML [Guindon and Gascuel, 2003]), "subtree pruning and re-grafting" removes a subtree from the whole tree and places it on another edge (used in RAxML [Stamatakis et al., 2005]), and "tree bisection and reconnection" splits the tree into two subtrees that are rewired by any of their edges. In all these rearrangements, whole subtrees remain fixed: their branches still have the same parameters and their internal topology is unchanged. By defining conditional likelihoods for fixed subtrees, and by placing the root at the rearrangement point, one can avoid much computation when exploring the space of tree topologies. As the root is placed at the rearrangement point, all the conditional likelihoods can be considered as lower from a mathematical point of view since none contains the root.

The most efficient algorithms first compute conditional likelihoods for all subtrees, before they compute an approximate likelihood for topologies obtained with a given sort of rearrangement, using the previously obtained conditional likelihoods. They apply the most promising rearrangements, either all at once (PhyML) or as soon as it is tried (RAxML), optimize evolutionary parameters of the new tree, and eventually start a new round of conditional likelihood calculation and exploration of the space of tree topologies, until convergence.

Computing the Likelihood of a Tree with a Nonreversible Model

Upper conditional likelihoods.—In the nonreversible case, upper conditional likelihoods can be defined to account for the true root of the tree and the evolutionary directions of the branches.

We define the upper conditional likelihood at branch RU in the nonreversible case as:

$$L_{s,upp(RU)}(R=x) = P(R=x) \times \sum_{z\in\Omega}[P_{xz}(l_A, v_A)L_{s,low(RA)}(A=z)] \quad (6)$$

The underlying branches' upper likelihoods can also be defined recursively:

$$L_{s,upp(UB)}(U=y) = \sum_{x\in\Omega}[P_{xy}(l_U, v_U)L_{s,upp(RU)}(R=x)] \times \sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)] \quad (7)$$

The main difference lies in the incorporation of the root nucleotide frequencies in the definition of the upper conditional likelihoods. This way, the root is not moved around the topology, and the evolutionary direction is conserved.

We now prove that the expression of the tree likelihood is not changed when computed from other nodes of the tree using upper and lower likelihoods.

Recurrence.—We show that for any branch, say UB,

$$L_s = \sum_{y\in\Omega}[L_{s,upp(UB)}(U=y)] \times \sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)}(B=q)]$$

We initialize the recurrence with branch RU:

$$L_{s,RU} = \sum_{x\in\Omega}\Bigg\{L_{s,upp(RU)}(R=x) \sum_{y\in\Omega}[P_{xy}(l_U, v_U)L_{s,low(RU)}(U=y)]\Bigg\} \quad (8)$$

We expand it:

$$L_{s,RU} = \sum_{x\in\Omega}\Bigg(P(R=x)\sum_{z\in\Omega}[P_{xz}(l_A, v_A)L_{s,low(RA)}(A=z)] \times \sum_{y\in\Omega}\Bigg\{P_{xy}(l_U, v_U)\sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)} (B=q)]\sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)]\Bigg\}\Bigg)$$

With (1):

$$L_{s,RU} = L_s$$

The likelihood computed on the branch *RU* is the same as the one computed at the root. This is also true for the edge *RA*.

We now suppose we know the likelihood at a branch *RU* and are interested in the likelihood at an underlying branch *UB*.

$$L_{s,UB} = \sum_{y\in\Omega}\Big\{L_{s,upp(UB)}(U=y) \times \sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)}(B=q)]\Big\} \quad (9)$$

We expand it, using (7):

$$L_{s,UB} = \sum_{y\in\Omega}\Big\{\sum_{x\in\Omega}[L_{s,upp(RU)}(R=x)P_{xy}(l_U, v_U)] \times \sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)] \times \sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)}(B=q)]\Big\}$$

And then rearrange it:

$$L_{s,UB} = \sum_{x\in\Omega}\Big\{L_{s,upp(RU)}(R=x) \times \sum_{y\in\Omega}[P_{xy}(l_U, v_U)L_{s,low(RU)}(U=y)]\Big\}$$

So that we have proved, with (8):

$$L_{s,UB} = L_{s,RU} = L_s$$

By recurrence, we have shown that the likelihood value can be computed from any branch of the tree, which, as will be seen, is particularly useful when exploring the space of tree topologies.

Exploring the Space of Tree Topologies with a Nonreversible Model

Efficient heuristics explore the space of tree topologies by local rearrangements such as nearest neighbor interchanges (NNIs). In the nonreversible case, evolution proceeds from the root of the tree to its leaves, so one must keep this evolutionary direction unchanged. For this purpose, a distinction is made between the branch on which the root is placed and the others. Figure 2a shows that being able to compute the likelihood from branch *UB* permits to define four subtrees whose conditional likelihoods can be used as constants to estimate the likelihoods of the three alternate topologies. In the nonreversible case, one uses upper conditional likelihood for the root-containing subtree and lower likelihoods for all other subtrees. With no loss of generality, NNIs around branch *UB* only require exchanges of subtrees having lower conditional likelihoods.

The likelihood of topology 1 Figure 2a can be computed with:

$$L_{s,a1} = \sum_{x\in\Omega}\Big\{L_{s,upp(RU)}(R=x)\sum_{y\in\Omega}\Big[(P_{xy}(l_U, v_U) \times \sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)] \times \sum_{q\in\Omega}\Big\{P_{yq}(l_B, v_B)\sum_{t\in\Omega}[P_{qt}(l_D, v_D)L_{s,low(BD)}(D=t)] \times \sum_{w\in\Omega}[P_{qw}(l_E, v_E)L_{s,low(BE)}(E=w)]\Big\}\Big]\Big\}$$

The likelihoods of topologies 2 and 3 are computed similarly.

In case the internal branch around which NNIs are to be done possesses the root, the situation is slightly different (Fig. 2b). As in the above case, three configurations can be reached through interchanges between subtrees, but here all conditional likelihoods are lower.

The likelihood of the topology 1 (Figure 2b) can be computed as follows:

$$L_{s,b1} = \sum_{x\in\Omega}\Big(P(R=x)\sum_{z\in\Omega}\Big\{P_{xz}(l_A, v_A) \times \sum_{i\in\Omega}[P_{zi}(l_F, v_F)L_{s,low(AF)}(F=i)] \times \sum_{j\in\Omega}[P_{zj}(l_G, v_G)L_{s,low(AG)}(G=j)]\Big\} \times \sum_{y\in\Omega}\Big\{P_{xy}(l_U, v_U)\sum_{q\in\Omega}[P_{yq}(l_B, v_B)L_{s,low(UB)} (B=q)]\sum_{v\in\Omega}[P_{yv}(l_C, v_C)L_{s,low(UC)}(C=v)]\Big\}\Big)$$

The likelihoods of topologies 2 and 3 are computed similarly.

It can be shown that by only doing NNIs, the root can be moved throughout the whole tree, to the exception of leaves: as NNIs keep internal branches internal (and external branches external), the root cannot be moved to a leaf.

Thus, the exploration through NNIs of the space of rooted tree topologies with a nonreversible model is as

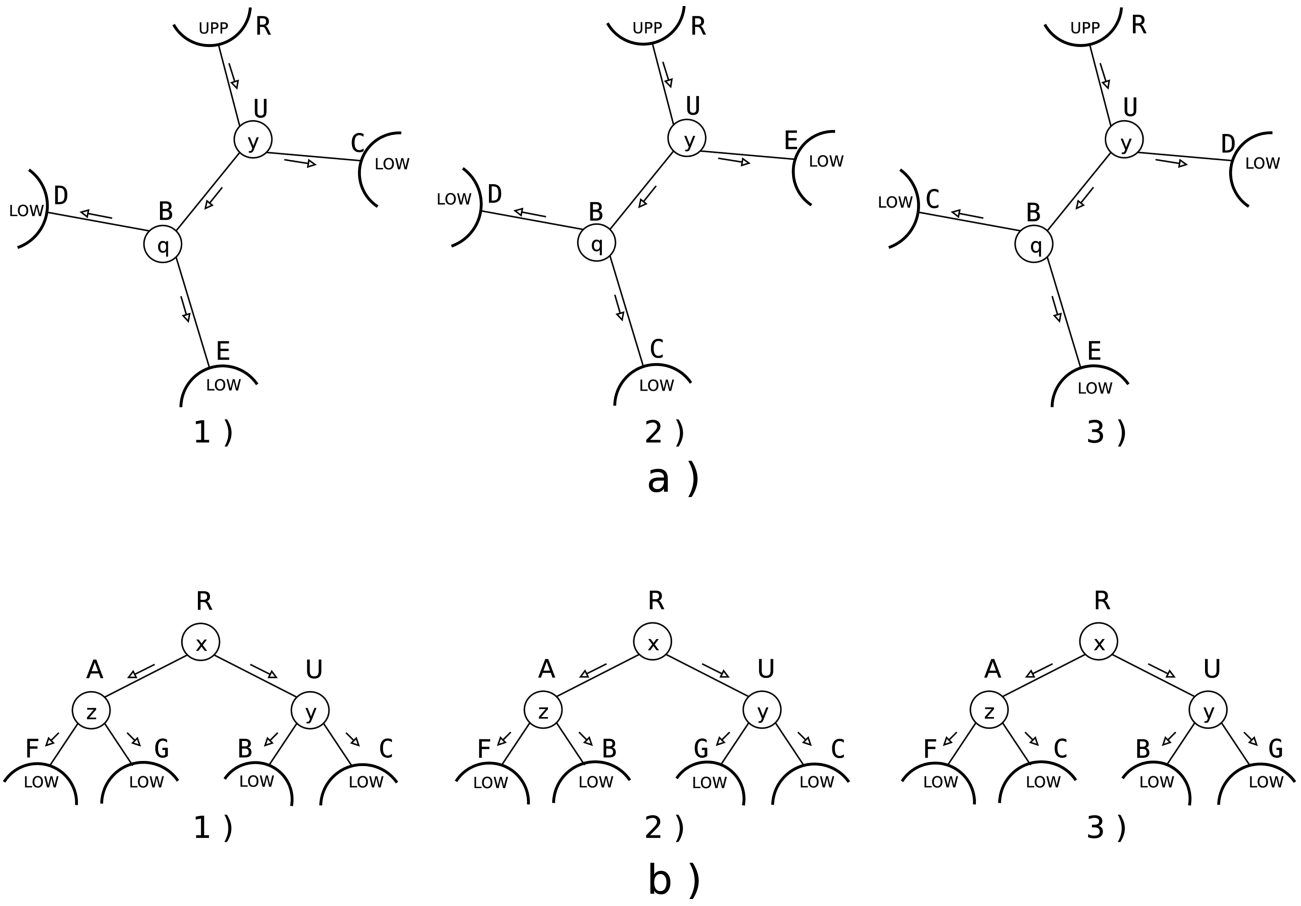


FIGURE 2. Use of conditional likelihoods when applying NNIs (Nearest Neighbor Interchanges) to an internal branch. (a) NNIs are applied to internal branch UB. The root node is situated in the subtree noted "UPP." The three other subtrees, named D, C, and E, are all lower. By swapping lower subtrees, three different topologies noted 1, 2, 3 are obtained. One can use the conditional likelihoods (upper in one case, lower in the three other cases) of the four subtrees to speed up the likelihood computation for these three alternate topologies. (b) The root node is now situated on the branch around which the NNI is done. All the conditional likelihoods are therefore lower, so any swap can be done.

exhaustive as the exploration of the space of unrooted trees with a reversible model, except that the root cannot go to an external branch.

Once again, and in the same way as computing the likelihood of a tree with a nonreversible model of evolution can be considered as a generalization of the reversible case, the way the space of tree topologies can be explored by NNIs with a nonreversible model of evolution can be seen as a generalization of the reversible case. Overall, nonreversible models of evolution can easily fit into recent heuristics such as PhyML to search for most likely rooted phylogenies.

In the next part, we report a new program built on the algorithmic architecture of PhyML, which explores the space of tree topologies under the nonhomogeneous, nonstationary model of Galtier and Gouy (1998).

nhPhyML, Adaptation of PhyML Algorithmic Structure to Galtier and Gouy's Model

We adapted the fast heuristics of PhyML (Guindon and Gascuel, 2003) to Galtier and Gouy's nonhomogeneous and nonstationary model, and we report here results concerning the ability of the resulting nhPhyML program to explore the space of tree topologies and to estimate the ancestral G+C content.

Galtier and Gouy's model is particularly variable rich: in addition to common parameters such as branch lengths and transition/transversion ratio, it incorporates different equilibrium G+C contents for each branch and an additional G+C content at the root. This makes it a model containing $4n - 2$ variables, with n the number of taxa in the tree: $2n - 3$ branch lengths, $2n - 2$ equilibrium G+C contents, the G+C content at the root, the transition/transversion ratio, and an additional parameter defining the position of the root on its branch; i.e., the fraction of the branch length lying on the left side of the root. All these parameters are estimated in the maximum likelihood framework, which leads to a computationally intensive model. For this reason, in all the studies that used this model to find phylogenetic trees (Galtier et al., 1999; Tarrio et al., 2001; Herbeck et al., 2005), no exploration of the space of tree topologies was conducted; the model was simply used to compare a limited set of input phylogenies.

The use of very efficient heuristics to find most likely trees is then mandatory for this kind of model to be used not just as an evaluation tool. PhyML (Guindon and Gascuel, 2003) is such an algorithm that explores the space of tree topologies around an input phylogeny. The adaptation of Galtier and Gouy's model to the algorithmic structure of PhyML permits for the first time to use this nonstationary, nonhomogeneous model to explore the space of phylogenies with dozens of sequences.

PhyML's code was deeply modified to produce nhPhyML. nhPhyML starts from a user input–rooted topology: the choice of the root depends upon the user and is unchanged throughout the whole search for the most likely tree, except for its position on its branch. Even if we have shown that NNIs around the root branch could be easily implemented, this has not been done in nhPhyML. Data structures had to be slightly remodeled, as in Galtier and Gouy's model each branch has its own substitution matrix, and algorithms had to incorporate the fact that the root was fixed, both in the computation of likelihood values for alternate topologies obtained by NNIs (2) and in the computation of conditional likelihoods themselves. Those conditional likelihoods are computed almost as in PhyML, by first a postorder (the original Felsenstein's pruning algorithm; Felsenstein, 1981) and then a preorder tree traversal, but starting from the root of the tree, whereas in the reversible case the starting point could be any leaf.

Equations (2) and (3) show that lower conditional likelihoods depend, if at a leaf, upon the base observed in the sequence, or, if at an internal node, upon the values of the underlying nodes lower conditional likelihoods. Lower conditional likelihoods can then be obtained by a postorder tree traversal starting from the root node: the tree is traversed to its leaves, and then the conditional likelihoods of the upper nodes are computed, from the leaves up to the root. On the contrary, upper conditional likelihoods depend both on underlying nodes' lower conditional likelihoods and on above-lying nodes' upper conditional likelihoods (Equations (6) and (7)): upper conditional likelihoods can then be computed once lower conditional likelihoods have been computed, and with a preorder tree traversal.

All the parameters of the model are optimized with the Newton-Raphson method (Felsenstein and Churchill, 1996; Galtier and Gouy, 1998). Derivatives are computed analytically except for the shape parameter of the gamma distribution accounting for differences in substitution rates across sites (Yang, 1993) whose derivatives are computed numerically.

The topology is reorganized as in PhyML except that when estimating the approximate likelihood of a given NNI, not only the length but also the equilibrium G+C content of the internal branch around which NNIs are done are optimized.

Results obtained with nhPhyML suggested that the program was prone to getting trapped in local maxima. This prompted us to develop an approximate version of the Galtier and Gouy model, named nhPhyML-Discrete.

We thus adopted a strategy inspired by Foster (2004) and only allowed a limited set of c equilibrium frequencies, themselves permanently set to $\frac{1}{c+1}, \ldots, \frac{c}{c+1}$. Each branch can still have its own equilibrium frequency, by choosing from the few ones available.

Three changes were introduced in the algorithm. First, the user sets the number c of equilibrium G+C frequencies. Second, before the exploration of the space of tree topologies, each equilibrium frequency is tested for each branch independently from the others, and the branch length is optimized for each equilibrium frequency. The best pairs (equilibrium frequency–branch length) are recorded and ordered according to the gain in likelihood they permit. When all branches have been tried, all the best values are simultaneously used. If the likelihood does not increase, only the first half of them, according to the order previously defined, are applied, until increase. This technique is very similar to the one used in PhyML to optimize branch lengths. Third, the space of tree topologies is explored as in PhyML, except that for each NNI that is tried, all the equilibrium frequencies are tried on the internal branch, and for each one the branch length is optimized.

Ability of nhPhyML to Explore the Space of Tree Topologies

In order to estimate the ability of nhPhyML to explore the tree topological space, we simulated the evolution of 1000-bp-long sequences according to Galtier and Gouy's model with a gamma-distributed rate across sites and the rooted version of the trees containing 40 leaves that were used to test PhyML (Guindon and Gascuel, 2003). The ancestral sequence G+C content was uniformly drawn from the interval [0.2; 0.8], and the equilibrium G+C frequencies were uniformly drawn at each node from the interval [0.1; 0.8], but the transition-transversion ratio was kept constant on the whole tree. We then applied various algorithms (neighbor joining, maximum parsimony and maximum likelihood with PhyML) to estimate their ability to find the topologies that had been used to simulate the evolution of the sequences (the "true topologies") and compared them to nhPhyML.

Among these algorithms, we distinguished programs that do not need a starting topology (like distance-based approaches) from the ones that reorganize a user input tree to explore the space of tree topologies (like nhPhyML). PhyML and the parsimony algorithm can be said to belong to the two classes, as they reorganize a starting topology that can be input a priori by the user or generated by the program itself. Two experiments were then conducted, one in which algorithms that do not need a user input topology were tested upon the simulated sequences (Fig. 3, white bars), and one in which algorithms that can run starting from a user input topology were compared (Fig. 3, grey bars). For this second experiment, input topologies were obtained by perturbing the "true topologies" by a number of NNIs uniformly drawn from [5; 20], while making sure that the ingroup and the outgroup were not melted. This additional condition is necessary as nhPhyML is not able to question

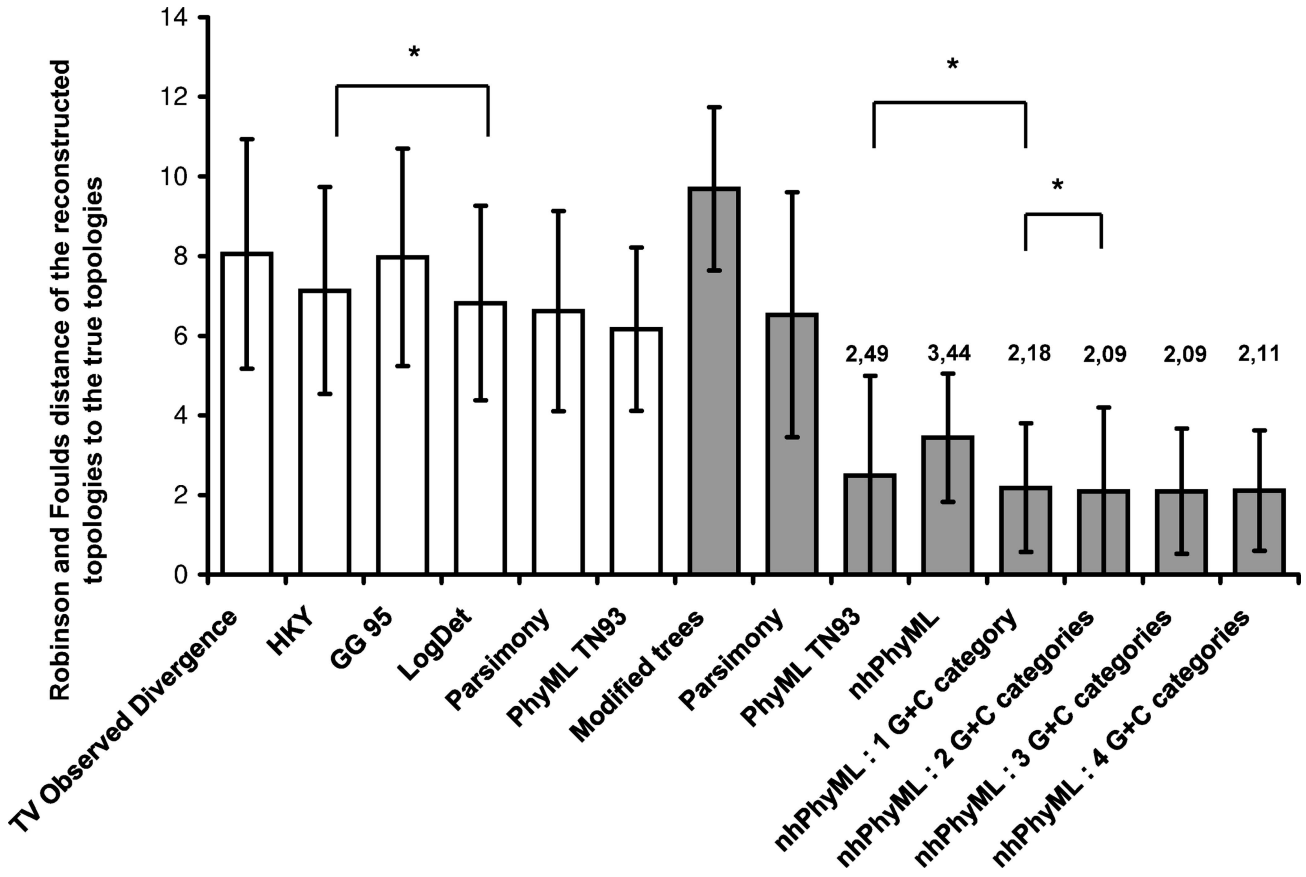


FIGURE 3. Efficiency of various methods in reconstructing a phylogeny in nonhomogeneous conditions. White bars: Results obtained by phylogenetic methods that do not start from a user input topology. Grey bars: Results obtained by phylogenetic methods that start from perturbed input topologies. These results were obtained with the same 2000 different topologies taken from the PhyML test set. Error bars represent standard deviations. Asterisks are displayed where Student paired *t*-tests are significant at the 1% level.

the position of the root when exploring the space of tree topologies.

To estimate the efficiency of a method, Robinson and Foulds' (R&F; Robinson and Foulds, 1979) average distances between the true topologies and the reconstructed ones were computed with the PHYLIP package (Felsenstein, 1989). Results are given in Figure 3.

For both PhyML (version 2.4.4, under the TN93 model [Tamura and Nei, 1993]) and nhPhyML, reconstruction was made using a gamma law with eight categories to account for across-site rate variation; parameter α, transition/transversion ratio, and the other parameters were estimated by the programs. PhyML and the parsimony method as implemented in PAUP* (Swofford, 2003) were both used from their built-in starting topology and from the perturbed input topologies. The neighbor-joining algorithm was applied to pairwise distances estimated under HKY85 (Hasegawa et al., 1985), LogDet (Lake, 1994; Lockhart et al., 1994), GG95 (Galtier and Gouy, 1995), and transversions-only observed divergence distances.

Figure 3 shows that PhyML is more efficient at finding good topologies than parsimony, which also has better results than distance methods. It is surprising to note that the transversions-only observed divergence and the GG95 distances perform worse than the HKY85 distance, because these methods were devised to be resistant to G+C content biases. However, as expected, the LogDet distance provides better results than the HKY85 distance.

Figure 3 (grey bars) compares the efficiencies of parsimony, PhyML, and nhPhyML methods. The average distance between the rearranged input topologies and the true topologies is shown. All the methods tested are able to explore the space of tree topologies to find better trees than the input ones. Results obtained by PhyML from the rearranged input phylogenies are better than when PhyML departs from its own starting topology (a distance-based tree). As the rearranged topologies are further away from the true topologies than the distance-based trees, this comes from the fact that distance trees fail on subtrees also difficult to solve for the maximum likelihood method, whereas the rearranged topologies can be perturbed at the level of subtrees whose solution is trivial.

It appears that PhyML is able to find better trees than parsimony and surprisingly also better trees than

nhPhyML (Student unilateral paired t-test, P-value $<1 \times 10^{-10}$). This might be due to the large number of parameters: nhPhyML has $2 \times n - 2$ additional parameters when compared to PhyML, because an equilibrium G+C content is associated to each branch. This might result in a likelihood surface with lots of local maxima, in which the algorithm would get trapped.

The comparison of the likelihood values found by nhPhyML when launched from the rearranged topologies to the likelihood values computed on the true topologies comforts us in this hypothesis: the log-likelihood of the true trees is on average 62.9 points higher than the log-likelihood of the trees found by nhPhyML, which have a better likelihood than the true trees in only 830 cases out of 2000. As a comparison, PhyML finds topologies with a 1.2-point higher log-likelihood score than the true topologies and finds topologies with better likelihoods than the true ones in 1562 cases out of 2000. This means that nhPhyML fails to correctly explore the space of tree topologies, because it gets trapped in local maxima and does not get to the real maximum. Being particularly parameter rich, it seems that nhPhyML can fit nearly any topology by taking advantage of its numerous parameters.

An approximate and less flexible model was developed and was named nhPhyML-Discrete. This nonhomogeneous model still has the same number of free parameters as nhPhyML, as each branch can have a particular equilibrium frequency, but is also much less "flexible," because these equilibrium frequencies are constrained to be in the limited set defined by the user. These constraints have a positive impact on the results of the algorithm. These are shown in Figure 3 for sets of 1 to 4 equilibrium frequencies.

nhPhyML-Discrete shows a better topological accuracy than PhyML even when using only one equilibrium frequency: this can be due either to the fact that the root base distribution is estimated in nhPhyML-Discrete and not in PhyML, or to the fact that the ingroup and the outgroup cannot be swapped in nhPhyML-Discrete, thereby avoiding the exploration of unreasonable tree topologies contrary to PhyML. Increasing the number of equilibrium frequencies further increases nhPhyML-Discrete's topological accuracy, but this tendency quickly reverses, as using 3 equilibrium frequencies yields better results than 4 equilibrium frequencies (though the unilateral paired Student t-test is not significant). When further increasing the number of equilibrium frequencies, the topological accuracy continues dropping, with, for instance, an average distance to the true topologies of 2.23 for 10 equilibrium frequencies (data not shown), not better than when using only 1 equilibrium frequency (2.18, Student unilateral paired t-test P-value: 0.054). Overall, it seems that using 3 equilibrium frequencies might be a good choice, as it gets the best topological accuracy on the simulations, which is the same performance as with 2 equilibrium frequencies, but with a lower standard deviation.

When used with 3 G+C equilibrium frequencies, the algorithm finds topologies closer to the true ones than PhyML (nhPhyML-Discrete: 2.09, PhyML: 2.49, Student unilateral paired t-test, P-value $<1 \times 10^{-10}$). This also has an impact on the risk of getting trapped in local optima: nhPhyML-Discrete finds topologies that have a log-likelihood on average 37.5 points higher than the true topology log-likelihoods, which is better than nhPhyML. Moreover, it appears that nhPhyML-Discrete finds trees with better likelihoods than the true tree in 1768 cases out of 2000. Overall, it seems that nhPhyML-Discrete shows a performance as good as PhyML's one, with a stronger variation in log-likelihood, which might hint for a stronger discriminating power. Finally, this approximation has a great impact on the computational speed, nhPhyML used on average 40 min 42 s to give its results while nhPhyML-Discrete only needed 20 min 43 s: it is faster to choose among a limited set of equilibrium frequencies than to optimize the value of a continuous parameter.

Figure 4 shows that nhPhyML-Discrete is, as PhyML and parsimony, nearly insensitive to the distance of its input phylogeny to the true one. On the contrary, nhPhyML does not seem to be able to cope with distant topologies, which is in agreement with the results above.

Estimation of Root G+C Content

The ancestral G+C content is a parameter of the model in itself. We conducted tests to check the ability of the program to estimate this parameter on trees containing 40 leaves, either from the true phylogenies or from the phylogenies found by nhPhyML and nhPhyML-Discrete when it was input perturbed topologies. nhPhyML-Discrete results are shown Figure 5 for 3 equilibrium frequencies.

Whether it is estimated from the true phylogenies, or from the phylogenies found by nhPhyML or nhPhyML-Discrete (Fig. 5), the ancestral G+C content is well estimated. The correlation coefficient between estimated and expected G+C contents is above 0.99 in all cases. Interestingly, results do not depend upon the number of equilibrium frequencies: performances are highly similar whether we use only 1 equilibrium frequency or whether we use nhPhyML. The average of the squared differences between the estimated and the true values is $\approx$0.000282 when inferred from the true phylogeny, $\approx$0.000292 when inferred from the phylogenies found by nhPhyML, and $\approx$0.0003423 when inferred by nhPhyML-Discrete with 3 equilibrium frequencies from the phylogenies it has found. It is interesting to note that even if nhPhyML-Discrete does not model evolution as precisely as nhPhyML, being limited in its choice of equilibrium frequencies, it can still provide very good estimates of the ancestral G+C contents, even from topologies that are not the true ones.

Overall it appears that the limitation of the number and values of equilibrium frequencies has been very successful, permitting to increase the ability of the algorithm to explore the space of tree topologies while retaining the capacity to estimate ancestral G+C content.

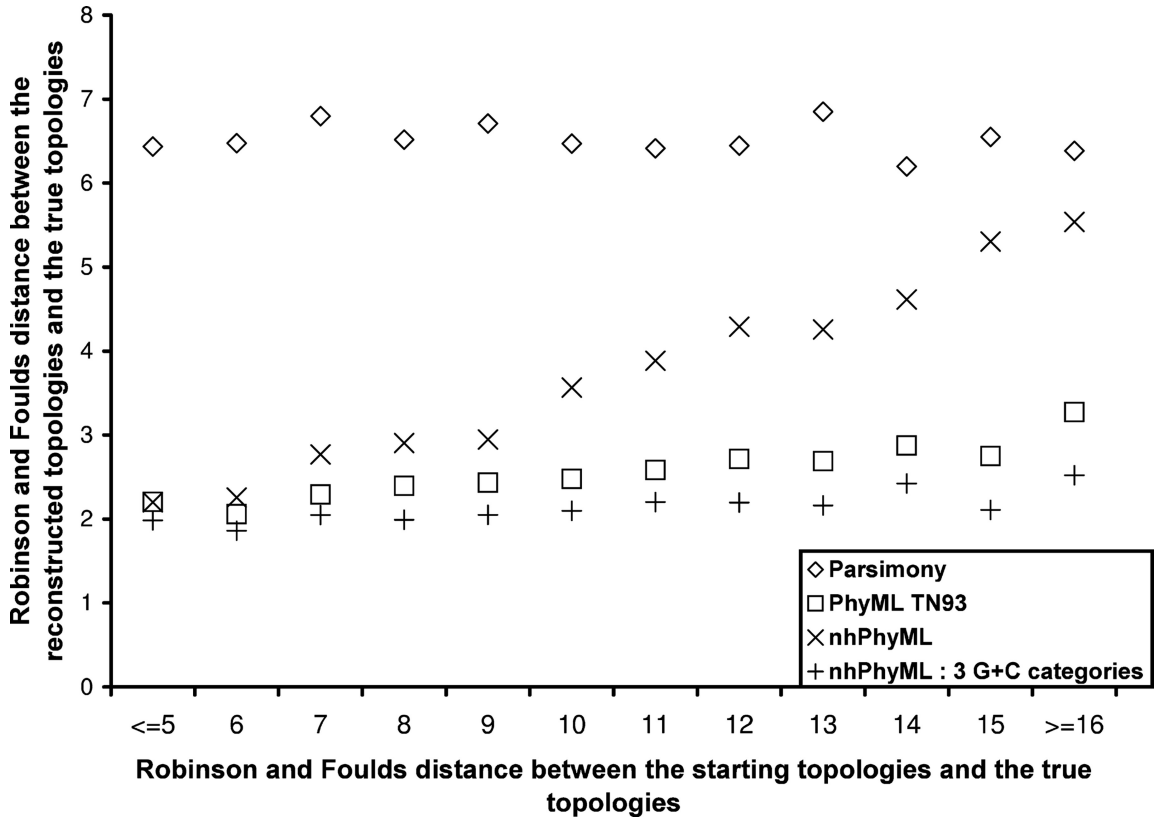


FIGURE 4. Ability of phylogenetic methods to explore the space of tree topologies in nonhomogeneous conditions. As the distance from the input phylogenies to the true topologies increases, results obtained by nhPhyML get worse, but nhPhyML-Discrete seems to be less dependent upon the input topologies.

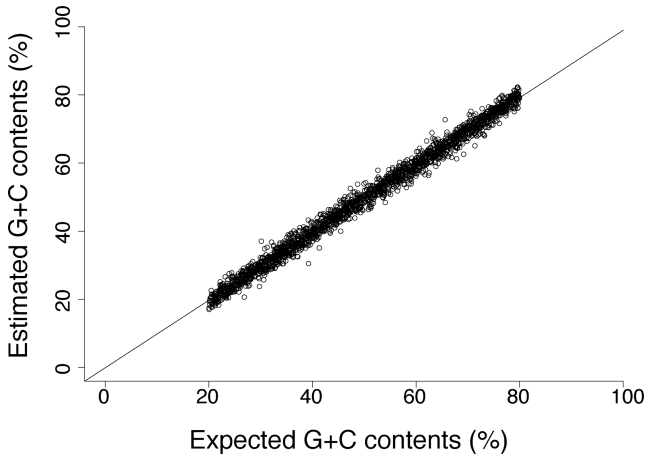


FIGURE 5. Ability of nhPhyML-Discrete to estimate ancestral G+C contents. nhPhyML-Discrete was used to estimate ancestral G+C contents on sequences simulated on 2000 different topologies from the PhyML test set (see text) from the phylogenies found by the program itself, with 3 equilibrium frequency categories. In each experiment, parameter α that accounts for across-site rate variation using an 8-category discretized gamma distribution, transition/transversion ratio, and the other parameters were estimated by maximizing the likelihood.

ESTIMATION OF THE UNIVERSAL PHYLOGENY AND OF THE ANCESTRAL G+C CONTENT

The ability of nhPhyML-Discrete to explore the space of tree topologies and to estimate the ancestral G+C content has then been demonstrated and can be applied to real data.

As rRNA stem G+C contents are known to be correlated to the optimal growth temperatures in Bacteria and Archaea (Galtier and Lobry, 1997), these genes are especially good candidates for an analysis using Galtier and Gouy's model (Galtier and Gouy, 1998), and hence nhPhyML-Discrete. Therefore, we reiterated the analysis of Galtier et al. (1999), improving the taxonomic sampling and benefiting from the heuristics of PhyML to better scan the space of tree topologies.

The analysis was divided in two steps: first we searched with nhPhyML-Discrete for the most likely topology for which complete rRNA genes and the model plead, and then, using the best topology we could find and the stem portion of rRNA sequences, we estimated the G+C content of the last universal common ancestor (LUCA) with nhPhyML.

Phylogeny Estimation

Small and large rRNA subunit sequences were downloaded from the European Ribosomal RNA database

(Wuyts et al., 2004) and from generalist databases for a few missing sequences. Ninety-two species were selected, representing the three domains of life with 22 Archaea, 34 Bacteria, and 36 Eukaryota. Small- and large-subunit rRNA genes were concatenated, aligned using ClustalW (Chenna et al., 2003), and manually curated. The resulting alignment contained 2924 sites, with G+C contents ranging from 43% to 71%, which highlights the need for models robust to compositional biases.

Even if the ability of nhPhyML-Discrete to explore the space of tree topologies appears as good as PhyML's, it is wise to run the program from various starting topologies to diminish the risk of getting trapped in local optima. Moreover, as the process of evolution in not reversible, the position of the root influences the likelihood value. For this reason, it was decided to build various topologies with distance-based and parsimony methods (as implemented in Phylo_Win, Galtier et al., 1996), and then to try three different rootings for each of these topologies. The trees were rooted either on the branch leading to the Archaea, on the branch leading to the Bacteria, or on the branch leading to the Eukaryota. This produced 57 starting phylogenies, among which 9 had identical topologies.

Results obtained by nhPhyML-Discrete on these trees were then analyzed using CONSEL (Shimodaira and Hasegawa, 2001) and Treedist (PHYLIP package; Felsenstein, 1989). Eighteen phylogenies were found to be significantly more likely than the others (AU test P-values <5%) among the 55 different topologies found, which shows that even if nhPhyML-Discrete showed good capabilities to explore the space of tree topologies, on real cases, with many sequences, the algorithm can get trapped in a local maximum, even when starting from two phylogenies with the same topologies but different branch lengths. None of these 18 topologies were identical, so the majority rule (extended) consensus tree obtained from these trees was built (Fig. 6) using Consense from the PHYLIP package (Felsenstein, 1989).

The Tree of Life

Though we do not believe that a two-gene phylogeny can clarify the Tree of Life, we think the analysis of rRNAs using a nonhomogeneous model might bear some interesting insights.

Most great clades are found monophyletic; e.g., Proteobacteria, Metazoa, Crenarchaeota. In the Bacteria kingdom, Firmicutes and Clostridia are found associated in all trees. Planctomycetes appear monophyletic and are associated to Chlamydiales but do not get a basal position as in Brochier and Philippe (2002): this result was found by getting rid of fastest evolving positions on the small subunit rRNA gene and is not found when coping with compositional heterogeneities on LSU and SSU rRNA genes. Instead, hyperthermophilic Bacteria are found at the root of the bacterial clade, as in first rRNA phylogenies (Woese, 1987).

Ancestral G+C Content Estimate

The consensus universal phylogeny found using nhPhyML-Discrete was used to infer the ancestral G+C contents of the small- and large-subunit rRNA genes stem regions. The inference was performed using only slowly evolving stem regions for two reasons. First, only rRNA stems, that is, the fraction of the rRNA molecule that folds as double helices, have a G+C content strongly correlated with optimal growth temperature (Galtier and Lobry, 1997). Second, Gowri-Shankar and Rattray (2006) have recently shown that the ancestral G+C content estimate obtained with the Galtier and Gouy model was biased towards the G+C content at slowly evolving sites and that equilibrium frequencies were biased towards fast-evolving sites. Correlations between interacting sites of the helices were not taken into consideration.

Stem regions were identified using the following procedure. The rRNA alignment downloaded from the European Ribosomal RNA database (Wuyts et al., 2004) was extended to the present sample of 92 sequences by manually aligning missing sequences using Seaview Galtier et al., (1996). A total of 1896 sites were predicted in stems in *Escherichia coli*, *Archaeoglobus fulgidus*, and *Schizosaccharomyces pombe.* Slowly evolving sites were identified by using the COE program (Dutheil et al., 2005) from the Bio++ package (Dutheil et al., 2006) and selecting sites predicted to undergo on average less than 0.1 substitution per branch under a HKY85 model with an 8-class discretized gamma law to model rate heterogeneity. Six hundred seventy-eight slowly evolving stem sites were finally retained.

The correlation between rRNA stem slowly evolving sites G+C content and optimal growth temperature (Topt) is high for both Bacteria (0.815) and Archaea (0.953), which is in agreement with Galtier and Lobry (1997). Because ancestral G+C content inferences are known to be robust with respect to the tree topology (Galtier et al., 1999; and Fig. 5 herein), those estimates are expected not to depend strongly on the input phylogeny.

Because the location of the root of the universal tree is not currently known (Brown and Doolittle, 1997; Forterre and Philippe, 1999), the likelihoods of all three possible rootings, that is, on the branch leading to each one of the three domains, were computed using nhPhyML on the slowly evolving stem sites. We chose not to use the discrete version of nhPhyML in order to avoid any bias that might arise from the fact that equilibrium frequencies are set to a priori values. For instance, with three G+C categories, the equilibrium frequencies are set to 0.25, 0.50, or 0.75, which may not model the evolution of the slowly evolving stem sites appropriately, given that their G+C content range from 0.52 for *Entamoeba histolytica* to 0.83 for *Methanopyrus kandleri*. Four rate categories were used to model rate heterogeneity, and the parameter α of the gamma distribution and the transition/transversion ratio were optimized by maximizing the likelihood. Since studies (Huelsenbeck et al., 2002; Yap and Speed, 2005) have shown that there may be information in extant sequences for the identification of the root position of a phylogenetic tree using nonreversible models of evolution, we compared all three likelihoods to see which root the model and the data were predicting. We used the slowly evolving stem sites, as they are expected to be less prone to saturation problems. The

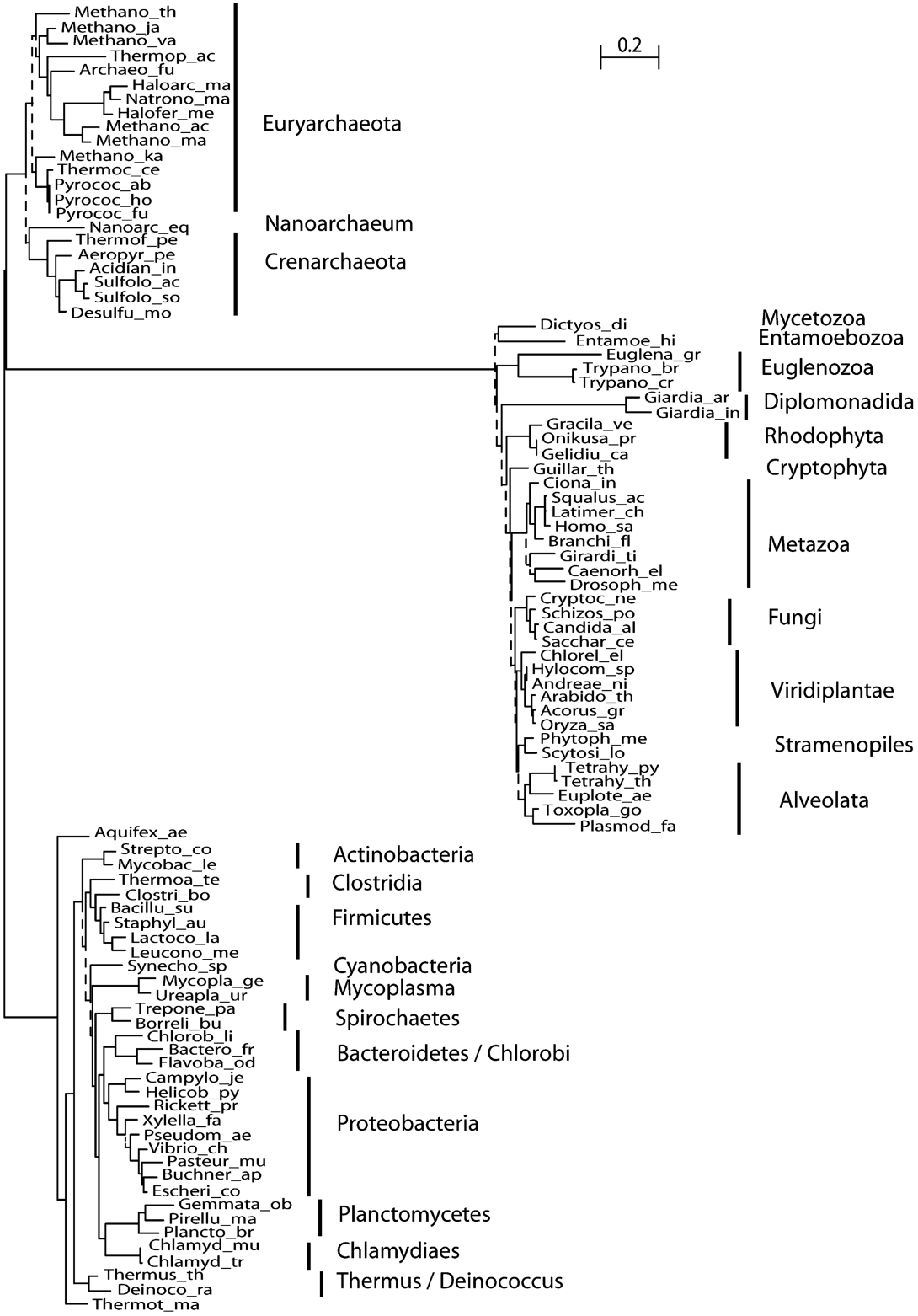


FIGURE 6. Consensus tree obtained from the 18 significantly most likely trees obtained by nhPhyML-Discrete. The tree was built as described in the text. Dashed lines represent branches that were not found in all the 18 most likely trees. The alignment contained 2924 sites. Group names are in agreement with the NCBI taxonomy. Sequence G+C contents range from 43% to 71%, which highlights the need to use models robust to compositional biases.

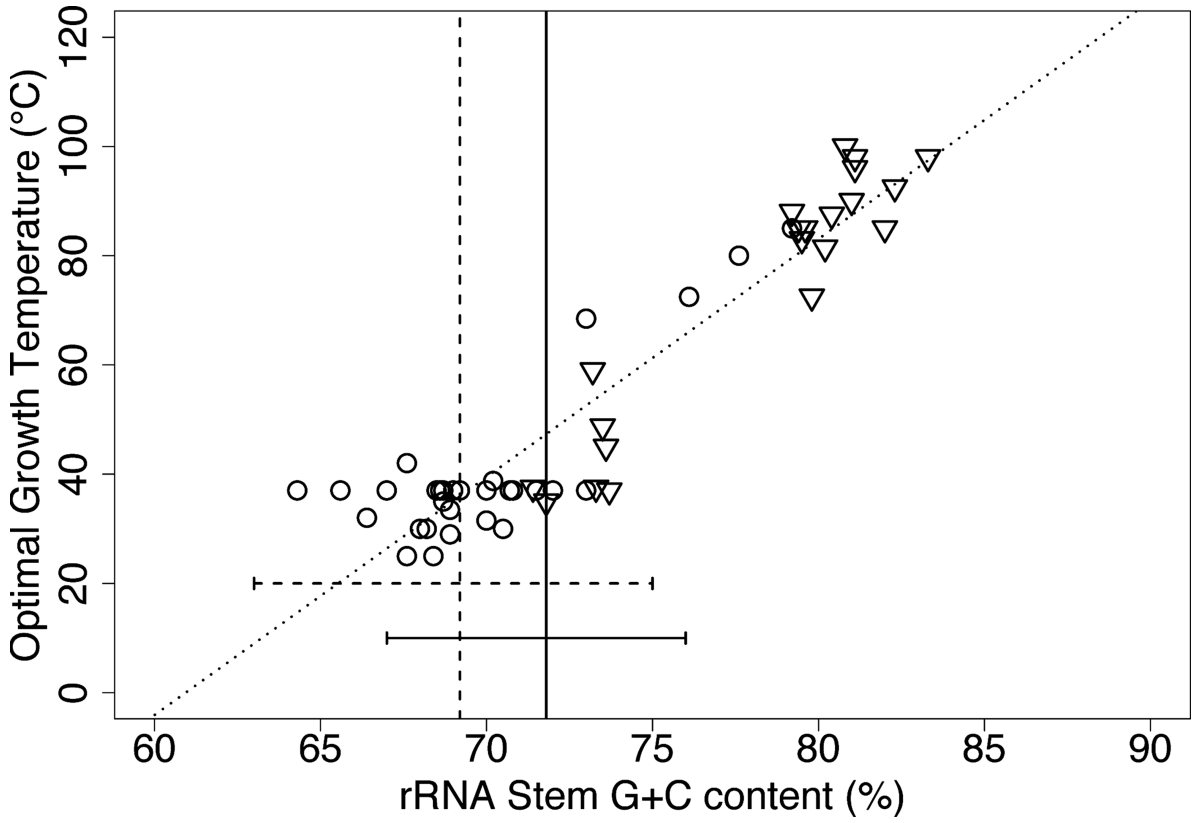


FIGURE 7. Correlation between rRNA stem G+C contents and optimal growth temperature, and ancestral G+C content estimates. Circles: bacterial data; triangles: archaeal data. Estimates of the G+C content of the stem fraction slowly evolving sites of rRNA molecules at the tree root are indicated by vertical lines, continuous for the bacterial branch rooting and dashed for the eukaryotic branch rooting. Confidence intervals are represented by horizontal segments, continuous for the bacterial rooting, and dashed for the eukaryotic rooting. The linear regression is represented as a dotted line. An orthogonal regression provided very similar results.

highest likelihood was obtained by the bacterial branch rooting (log-likelihood: −15,139.94, ancestral G+C content: 71.8%), whereas the least realistic rooting was found on the archaeal branch (log-likelihood: −15,147.22, ancestral G+C content: 75.1%) and could be rejected using CONSEL (Shimodaira and Hasegawa, 2001) (AU test P-value <5%). Hence, we chose to discard the archaeal rooting from subsequent analyses. There was no significant difference between the bacterial and the eukaryotic rootings (log-likelihood: −15,141.61, ancestral G+C content: 69.2%). It seems interesting to note that the longest branch (the eukaryotic branch) was not found to provide the most likely rooting: the model is then able to find a signal that is independent from the evolutionary distance.

To estimate the accuracy of the estimation of ancestral G+C contents, we computed the likelihoods obtained when setting the ancestral G+C content to various values, between 55% and 85%, and compared the results using CONSEL. Ancestral G+C contents with AU test P-values higher than 5% were considered to be in the confidence interval. We found that, when rooting in the eukaryotic branch, ancestral G+C contents ranging from 63% to 75% could not be rejected, whereas when rooting in the bacterial branch, the confidence interval was [67%; 76%]. The larger confidence interval found for the eukaryotic rooting might be explained by the fact that this branch is considerably longer than the bacterial one: the extra length of the eukaryotic branch may provide more latitude to accomodate nonoptimal ancestral G+C contents.

Inferred ancestral G+C contents (Fig. 7) suggest a mesophilic (optimal growth temperature below 60°C) to thermophilic (optimal growth temperature between 60°C and 80°C) LUCA, in agreement with Galtier et al. (1999). Interestingly, both confidence intervals do not contain any value that seem to favour a hyperthermophilic LUCA. As a consequence, it appears that reducing site rate heterogeneity to avoid the bias put forth by Gowri-Shankar and Rattray (2006) does not contradict Galtier et al.'s conclusion.

CONCLUSION

In this article, we have shown that by reorganizing the way likelihood is computed, one can efficiently explore the space of tree topologies with a nonreversible model of evolution. We modified the PhyML algorithm

(Guindon and Gascuel, 2003) to cope with the nonhomogeneous, nonstationary model of Galtier and Gouy (1998) and tested its abilities to find the right topology and to estimate the G+C content at the root. An approximate model was also tested, which showed good performance in both tree topological space exploration and ancestral G+C content estimation. We eventually used the program to estimate the topology of the Tree of Life from rRNA sequences and to estimate the ancestral stem G+C content by only selecting slowly evolving sites. The results agree with the ones obtained by Galtier and Lobry (1999) and support a nonhyperthermophilic last universal common ancestor.

Genome sequences vary widely in their composition between species. Therefore, when building a phylogenetic tree from such heterogeneous data, it is important to use a method robust to compositional biases. Nonhomogeneous models of evolution are particularly suitable, but their nonreversibility discarded them from most general phylogeny packages and prevented their use in large scale analyses. This work renders nonreversible models of evolution useful for phylogeny reconstruction, which considerably broadens the range of available models and opens new opportunities for models explicitly dealing with compositional biases.

ACKNOWLEDGEMENTS

nhPhyML is available as a LINUX executable file at http://pbil.univ-lyon1.fr/software/nhphyml/ and as source code upon request. We want to thank the reviewers and Olivier Gascuel for their constructive remarks, which resulted in considerable improvement in the quality of the manuscript. This work was supported by Action Concertée Incitative IMPBIO. We thank the Centre de Calcul de l'IN2P3 for providing computer resources. Bastien Boussau acknowledges a PhD scholarship from the Centre National de la Recherche Scientifique. We also thank Mathilde Paris for fruitful discussions.

REFERENCES

Brochier, C., and H. Philippe. 2002. Phylogeny: A non-hyperthermophilic ancestor for bacteria. Nature 417:244–244.

Brown, J. R., and W. F. Doolittle. 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol Mol. Biol. Rev. 61:456–502.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson. 2003 Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31:3497–3500.

Dutheil, J., T. Pupko, A. Jean-Marie, and N. Galtier. 2005. A model-based approach for detecting coevolving positions in a molecule. Mol. Biol. Evol. 22:1919–1928.

Dutheil, U., U. Gaillard, U. Bazin, U. Glemin, U. Ranwez, U. Galtier, and U. Belkhir. 2006. Bio++: A set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. BMC Bioinformatics 7:188–188.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 1989. Phylogeny inference package (version 3.2). Cladistics 5:164–166.

Felsenstein, J., and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93–104.

Forterre, P., and H. Philippe. 1999. Where is the root of the universal tree of life? Bioessays 21:871–879.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Galtier, N., and M. Gouy. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA 92:11317–11321.

Galtier, N., and M. Gouy. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871–879.

Galtier, N., M. Gouy, and C. Gautier. 1996 SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. Cabios 12:543–548.

Galtier, N., and J. R. Lobry. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. 44:632–636.

Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221.

Gowri-Shankar, V., and M. Rattray. 2006. On the correlation between composition and site-specific evolutionary rate: Implications for phylogenetic inference. Mol. Biol. Evol. 23:352–364.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Herbeck, J. T., P. H. Degnan, and J. J. Wernegreen. 2005. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria). Mol. Biol. Evol. 22:520–532.

Huelsenbeck, J. P., J. P. Bollback, and A. M. Levine. 2002. Inferring the root of a phylogenetic tree. Syst. Biol. 51:32–43.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. USA 91:1455–1459.

Lockhart, P. J., M. A. Steel, M. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence. Mol. Biol. Evol. 11:605–612.

Robinson, D., and L. Foulds. 1979. Comparison of weighted labeled trees. Pages 119–126 *in* Isomorphic factorisations VI: Automorphisms, combinatorial mathematics (A. F. Horadam and W. D. Wallis, eds.). No. 748 in Lecture Notes in Mathematics, Springer, Berlin.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.

Swofford, D. L. PAUP*. 2003. Phylogenetic analysis using parsimony (*and other methods), version 4. Sinauer Associates, Sunderland, Massachusetts.

Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. 9:678–687.

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. 10:512–526.

Tarrio, R., F. Rodriguez-Trelles, and F. J. Ayala. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. 18:1464–1473.

Woese, C. R. 1987. Bacterial evolution. Microbiol Rev. 51:221–271.

Wuyts, J., G. Perriére, and Y. Van De Peer. 2004. The European Ribosomal RNA database. Nucleic Acids Res. 32:D101–D103.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573.

Yap, V. B., and T. Speed. 2005. Rooting a phylogenetic tree with nonreversible substitution models. BMC Evol. Biol. Jan 4;5(1):2.

First submitted 17 October 2005; reviews returned 10 December 2005; final acceptance 28 March 2006.
Associate Editor: Olivier Gascuel

5

An Unexpected Archaea

Here, the precedently developed program nhPhyML was used to tackle another phylogenetic issue, that of the position of the archaea *Cenarchaeum symbiosum*, by analysing its rRNA sequences. The result did not permit to provide a firm answer, but further work by Céline Brochier-Armanet, Simonetta Gribaldo and Patrick Forterre led us to propose that *Cenarchaeum symbiosum* was so unlike other Archaea that it might deserve to be part of a new archaeal phylum, Thaumarchaeota.

When analysing a dataset containing more than 400 sequences, I was led to see that nhPhyML had some problems exploring the space of tree topologies. More work is required on the algorithmics associated with parameter-rich models.

This article has been published in *Nature Reviews Microbiology*.

Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota

Céline Brochier-Armanet, Bastien Boussau‡, Simonetta Gribaldo§ and Patrick Forterre§ ||*

Abstract | The archaeal domain is currently divided into two major phyla, the Euryarchaeota and Crenarchaeota. During the past few years, diverse groups of uncultivated mesophilic archaea have been discovered and affiliated with the Crenarchaeota. It was recently recognized that these archaea have a major role in geochemical cycles. Based on the first genome sequence of a crenarchaeote, *Cenarchaeum symbiosum*, we show that these mesophilic archaea are different from hyperthermophilic Crenarchaeota and branch deeper than was previously assumed. Our results indicate that *C. symbiosum* and its relatives are not Crenarchaeota, but should be considered as a third archaeal phylum, which we propose to name Thaumarchaeota (from the Greek '*thaumas*', meaning wonder).

Hyperthermophile
An organism that has an optimal growth temperature of at least 80°C.

Paraphyletic
A group of organisms or sequences that includes an ancestor and some, but not all, of its descendants.

**Université de Provence, Aix-Marseille I, CNRS, UPR 9043, Laboratoire de Chimie Bactérienne, Institut de Biologie Structurale et de Microbiologie, 13402 Marseille, France.*
‡Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 69622 Villeurbanne, France.
§Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE), Département de Microbiologie, Institut Pasteur, 75015 Paris, France.
||Université Paris-Sud, 91405 Orsay, France.
Correspondence to C.B.
e-mail: celine.brochier@ibsm.cnrs-mrs.fr
doi:10.1038/nrmicro1852

The RNA component of the small subunit of the ribosome (referred to here as SSU rRNA) has been the 'Rosetta stone' of modern evolutionary studies[1]. In particular, the discovery of the archaeal domain and establishment of the evolutionary relationships between archaeal species were based entirely on rRNA studies[2–5]. These analyses led to the proposal that the archaeal domain should be divided into two phyla, the Euryarchaeota (from the Greek '*euryos*', meaning diversity) and the Crenarchaeota (from the Greek '*crenos*', meaning spring or origin)[6]. At that time, the Euryarchaeota included a mixture of methanogens, extreme halophiles, thermoacidophiles and a few hyperthermophiles. By contrast, the Crenarchaeota included only hyperthermophiles (hence their name, which refers to a 'hot origin of life' hypothesis). This division of the Archaea was rapidly accepted, because it had been observed in the early days of archaeal research that Sulfolobales and Thermoproteales (two hyperthermophilic crenarchaeota orders) are fundamentally different to other archaea in terms of their SSU rRNA oligonucleotide catalogues[7] and RNA polymerase structures[8].

More recently, genomic data[9] and gene phylogenies that have been obtained from combined datasets[10–12] have also confirmed the division of the Archaea into two main lineages, although Euryarchaeota are sometimes paraphyletic in whole-genome trees, probably owing to artefacts that have been introduced by horizontal gene transfer (HGT) from bacteria[13,14]. Several genes that are involved in key cellular processes in the Euryarchaeota lack homologues in all hyperthermophilic crenarchaeota for which complete genome sequences are available[15–18]. For example, there are no homologues of the DNA polymerase from the D family[19] and the cell-division protein FtsZ[20] in hyperthermophilic crenarchaeota, both of which are present in all sequenced complete euryarchaeal genomes. Furthermore, this group of organisms lacks homologues of the eukaryotic-like histone[21] and the protein MinD (involved in chromosome and plasmid partitioning[15]), both of which are present in most sequenced euryarchaeal genomes. This indicates that important differences in main cellular processes were established shortly after the speciation of the Euryarchaeota and Crenarchaeota[14].

The discovery of mesophilic crenarchaeota

More than 20 years ago, direct PCR amplification of genes that encode the SSU rRNA from environmental samples gave rise to molecular ecology[22]. One of the major early outcomes of this new discipline was the discovery of many novel lineages of mesophilic or psychrophilic archaea[23,24] (reviewed in REFS 25,26). The first environmental archaeal sequences were detected in marine environments, and were clearly separated into two groups (named group I and group II) in an SSU rRNA tree that was rooted by a bacterial outgroup[23]. Group I formed a sister group of hyperthermophilic crenarchaeota, whereas group II emerged within the Euryarchaeota[23].

Possibly because they were discovered only 2 years after the generally accepted proposal to divide Archaea into 2 phyla[6], group I was classified as Crenarchaeota[23,24], even though it was only a sister group of hyperthermophilic crenarchaeota and did not branch off within them. The classification of group I archaea as Crenarchaeota was further strengthened by the phylogenetic analysis of a DNA polymerase sequence from *Cenarchaeum symbiosum* (a marine archaeon that inhabits the tissues of a temperate water sponge[27]), which branched within sequences from hyperthermophilic crenarchaeota[28]. Consistent with this, a recent, and widely accepted, SSU rRNA tree that was published by Schleper and colleagues[25,29], and has been widely used to illustrate archaeal phylogeny, shows mesophilic archaea of group I emerging within hyperthermophilic crenarchaeota. This phylogenetic placement is consistent with the assumption that mesophilic crenarchaeota evolved from hyperthermophilic ancestors through adaptation to a mesophilic lifestyle[11,14,30–32]. However, this placement remains controversial, because in most SSU rRNA phylogenies, such as the one recently published by Pace's group[33], group I sequences do not emerge within cultivated hyperthermophilic crenarchaeota and form a distinct lineage. Interestingly, the recent discovery of a eukaryotic-like histone gene that was probably not acquired by HGT in a genomic fragment from *C. symbiosum*[34] suggests that mesophilic crenarchaeota might have genomic features that are substantially different from those of hyperthermophilic crenarchaeota. Indeed, homologues of this gene are present in most euryarchaeal genomes, but never in hyperthermophilic crenarchaeota.

The ecological importance of mesophilic crenarchaeota, an extremely diverse group that is widely distributed in oceans and soils[35], is being increasingly recognized. Indeed, molecular environmental surveys have extended the diversity of mesophilic crenarchaeota by revealing several new lineages that are related to group I sequences, such as SAGMCG-1, FFS, marine benthic groups B and C, YNPFFA and THSC1 (reviewed in REFS 25,26). Some of these crenarchaeota might be moderate thermophiles or psychrophiles, even though the group is still designated as mesophilic crenarchaeota. Mesophilic crenarchaeota comprise organisms that are probably important participants in the global carbon and nitrogen cycles[25,36,37], and might be the most abundant ammonia oxidizers in soil ecosystems[37]. For example, it was reported that *Candidatus* Nitrosopumilus maritimus, a recently isolated mesophilic crenarchaeon, can grow chemolithoautotrophically by aerobically oxidizing ammonia to nitrite, which was the first observation of nitrification in the Archaea[38].

Investigating the phylogenetic position of mesophilic crenarchaeota within the archaeal phylogeny, together with their gene content and genomic features, could, therefore, provide valuable information on the evolution of the Archaea.

Can rRNA resolve deep archaeal phylogeny?

The phylogenetic position of mesophilic crenarchaeota is currently based solely on SSU rRNA sequences. The trees that were published by Schleper *et al.*[25] and Robertson *et al.*[33] included a large number of sequences (1,344 and 712 SSU rRNA sequences, respectively), but both showed poor resolution of the relative order of emergence of the different archaeal lineages and it was pointed out that the Crenarchaeota and Euryarchaeota appeared as polytomies (star radiations)[33]. This lack of resolution showed that SSU rRNA sequences do not contain enough phylogenetic signal to resolve the deepest nodes of the archaeal phylogeny, probably owing to their size, which limits the number of nucleotide positions that are available for phylogenetic analyses. However, the number of positions that can be used for phylogenetic analyses can be increased by a combined analysis of SSU and large subunit (LSU) rRNA sequences.

FIGURE 1 shows a maximum likelihood phylogenetic tree that is based on the concatenation of 226 SSU and LSU sequences from complete genomes that are representative of archaeal and bacterial diversity, as well as 18 mesophilic crenarchaeal or euryarchaeal fosmids that contain both types of sequences. Mesophilic crenarchaeal fosmid sequences belong to three distinct subgroups: groups 1.1a and 1.1b[25], and the recently proposed deep-branching HWCG III group[39]. The bacterial part of the tree shows a phylogeny that is consistent with those previously published (that is, high statistical support for the monophyly of most bacterial phyla, but a low resolution of their relative order of emergence (not shown)). For the Archaea, the monophyly of most orders within both Euryarchaeota and Crenarchaeota is robustly recovered (FIG. 1). However, the relationships among most euryarchaeal orders are poorly resolved (bootstrap value (BV) of less than 70%) (FIG. 1), and even the monophyly of the Euryarchaeota is not significantly supported (BV of less than 16%). Importantly, both mesophilic and hyperthermophilic crenarchaeota were recovered as two robust monophyletic groups (BV of 99 and 100%, respectively), which is consistent with the SSU rRNA tree published by Robertson and colleagues[33], but not with the tree that was published by Schleper and colleagues[25]. Moreover, mesophilic and hyperthermophilic crenarchaeota form a sister group, but with low support (BV of 36%), and the node is extremely unstable. For example, using a different evolutionary model, the position of mesophilic crenarchaeota was altered — they branched at the base of the archaeal tree and, therefore, became the sister group of a large group that included Euryarchaeota and hyperthermophilic crenarchaeota — but still with low statistical support (BV of 20%; not shown).

A possible explanation for such poor resolution could be the heterogeneity of G+C content among sequences. Sequences from hyperthermophilic euryarchaeota and crenarchaeota have higher G+C content compared with that of mesophilic organisms. This well-known compositional bias of RNA sequences might blur the genuine phylogenetic signal[40]. To investigate this possibility, we used a recently developed phylogenetic method that reduces the biases that are due to convergent G+C content (nhPHYML[41]). We tested three possible deep placements for mesophilic crenarchaeota, based on the rRNA archaeal phylogeny of FIG. 1: first, as a sister group of hyperthermophilic crenarchaeota; second, as a sister group of a cluster that comprises Euryarchaeota and

Sister groups
In a phylogeny, two lineages that share an exclusive common ancestor.

Monophyletic group
Includes an ancestor and all its descendants.

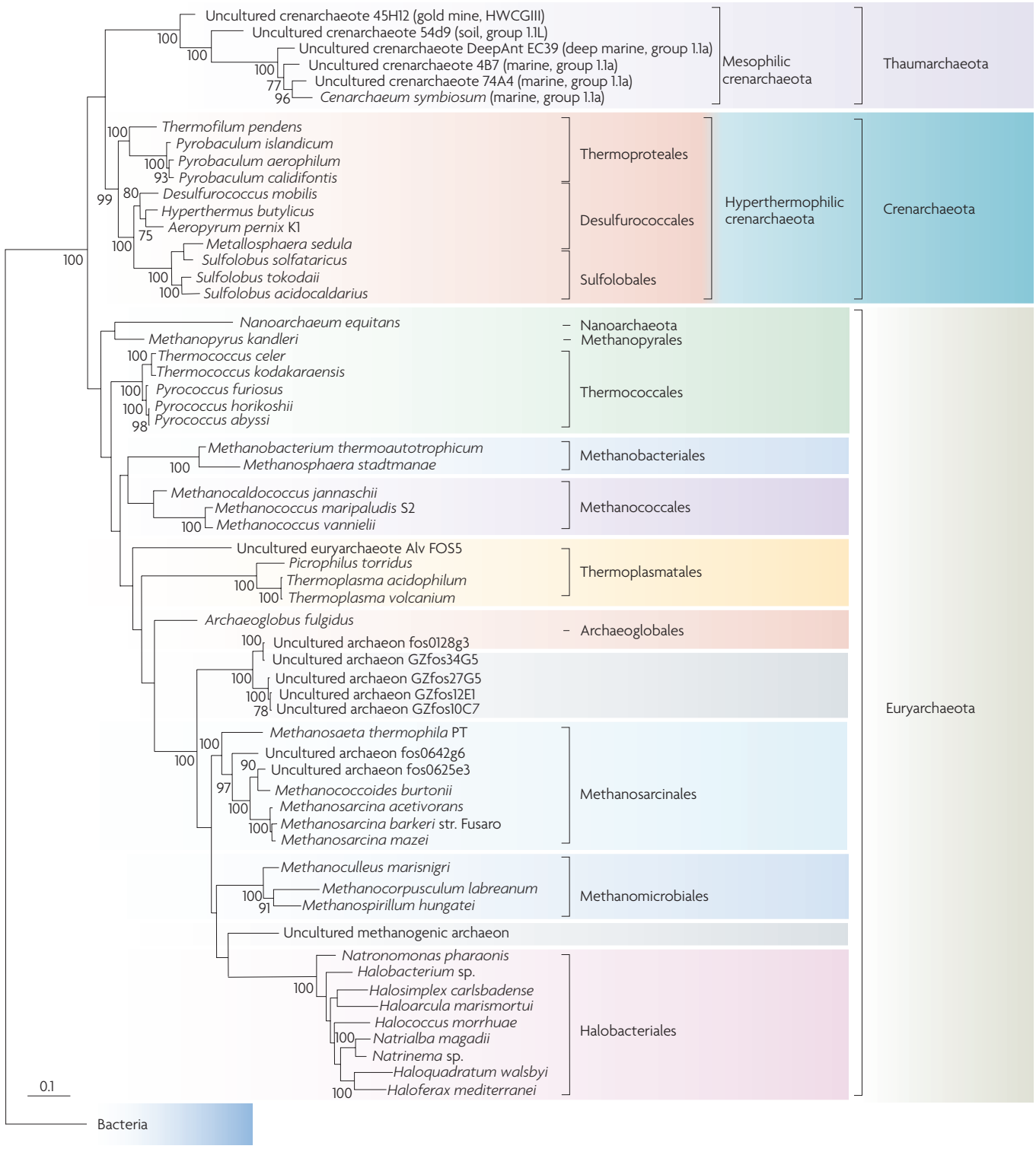


Figure 1 | **Maximum likelihood tree based on the concatenation of 226 SSU and LSU sequences from Archaea and Bacteria.** For clarity, the bacterial part of the tree is not shown. Sequences were aligned using MUSCLE (multiple sequence comparison by log-expectation)[58]. Resulting alignments were manually refined using the MUST (Management Utilities for Sequences and Trees) package[59], and only unambiguously aligned regions were kept for phylogenetic analyses. Concatenation was performed using home-developed software (C.B., unpublished data), which provided a final dataset of 3,305 nucleotide positions. The maximum likelihood tree was computed by PHYML[61], using the general time-reversible model of sequence evolution by including a Γ-correction (eight categories of evolutionary rates, an estimated α-parameter and an estimated proportion of invariant sites). Numbers at nodes represent non-parametric bootstrap values (BVs) that were computed by PHYML[61] (1,000 replications of the original dataset) using the same parameters. For clarity, only BVs of more than 70% are shown. The scale bar represents the average number of substitutions per site. If a different evolutionary model (Hasegawa Kishino Yano) was used, a sister grouping of hyperthermophilic crenarchaeota and euryarchaeota, and a basal branching of mesophilic crenarchaeota was recovered, albeit with weak statistical support (BV of 20%).

hyperthermophilic crenarchaeota; and, third, as a sister group of Euryarchaeota (Supplementary information S1 (table)). All six tests significantly rejected the third topology, whereas only two tests rejected the second topology. This means that the tests discard the third topology, but do not allow discarding the second topology in favour of the first topology. It is likely that the phylogenetic signal which is carried by rRNA sequences is too weak to confidently resolve the position of mesophilic crenarchaeota in the archaeal phylogeny, even if the number of positions is increased by combining SSU and LSU rRNA sequences. Nevertheless, the phylogenetic analysis of the rRNAs strongly supports the separation of mesophilic and hyperthermophilic crenarchaeota into 2 distinct lineages (BV of 100 and 99% for the monophyly of each lineage, respectively). To clarify the position of mesophilic crenarchaeota in the archaeal tree further, the use of alternative markers thus becomes crucial.

Analysing ribosomal proteins

Although they were first discovered 15 years ago, the isolation and cultivation of representative mesophilic crenarchaeota has proven to be a frustrating task. In fact, the first genome of a member of this group, *C. symbiosum,* which has still not been grown in pure culture, was published only recently[42]. The availability of this genome sequence now permits an investigation of the phylogenetic position of mesophilic crenarchaeota, based on markers other than SSU and LSU rRNA.

Owing to the availability of an increasing number of complete archaeal genomes, large concatenated datasets of ribosomal (R) proteins are now widely used as an alternative to SSU rRNA to study archaeal phylogeny[43–45]. Indeed, these proteins have the same evolutionary attributes as rRNA, and their concatenation allows the construction of larger alignments. Although the trees that were obtained using these markers were roughly congruent with the rRNA trees[43], they substantially improved the archaeal phylogeny and resolved a number of important nodes (reviewed in REFS 11,14). In particular, these analyses have helped to clarify the phylogenetic positions of 'lonely' archaeal species (those that lack sequenced relatives), which are often misplaced, especially if they are fast-evolving or have a biased sequence composition (for example, the G+C content of rRNA sequences)[46]. For example, *Nanoarchaeum equitans* was originally proposed to represent a third (and basal) archaeal phylum based on trees that were produced using SSU rRNA[47] and concatenated R proteins[44]. However, a subsequent analysis of R proteins and additional protein markers suggested that this species is not the earliest archaeal offshoot, but is probably a fast-evolving euryarchaeal lineage that is possibly related to Thermococcales[48]. Another example is the hyperthermophilic methanogen *Methanopyrus kandleri*, for which phylogenetic placement is crucial to obtain an understanding of the time of emergence of methanogenesis within Euryarchaeota. In fact, although *M. kandleri* represents the earliest euryarchaeal offshoot in SSU rRNA phylogenies[25,49], in trees that are based on R-protein concatenations it robustly branches off after the non-methanogenic lineage of Thermococcales[10,11,50]. Further, a recent phylogenetic analysis placed this archaeon as a sister group of two other methanogen lineages (Methanococcales and Methanobacteriales)[51], which is in agreement with phylogenomic studies of the genes that are involved in methanogenesis[51] and gene-content analyses[45]. Globally, these analyses indicate that methanogenesis might not be the ancestral metabolism of euryarchaeota.

The examples of *N. equitans* and *M. kandleri* highlight the power of R-protein combined datasets for phylogenetic reconstruction. We therefore applied the same approach to study the placement of *C. symbiosum* in the archaeal phylogeny. FIGURE 2 shows a maximum likelihood phylogeny of the archaeal domain that is based on the concatenation of 53 R-protein sequences from 48 complete archaeal genomes and was rooted using sequences from 16 eukaryotes. The phylogeny includes *C. symbiosum*, 33 Euryarchaeota and 14 hyperthermophilic crenarchaeota, which represents 21 new species (11 Euryarchaeota, and 1 mesophilic and 9 hyperthermophilic crenarchaeota, respectively) with respect to previous similar analyses[11]. This tree is better resolved than the SSU/LSU rRNA tree in FIG. 1 (note the higher BVs at nodes in FIG. 2), and the positions of the newly included archaea are well supported and in agreement with their classification. Consequently, *Thermofilum pendens*, *Caldivirga maquilingensis*, *Pyrobaculum calidifontis*, *Pyrobaculum arsenaticum* and *Pyrobaculum islandicum* are grouped with *Pyrobaculum aerophilum* (group of Thermoproteales; BV of 100%), and *Ignicoccus hospitalis*, *Staphylothermus marinus* and *Hyperthermus butylicus* are grouped with *Aeropyrum pernix* (group of Desulfurococcales; BV of 97%), whereas *Metallosphaera sedula* is grouped with other Sulfolobales (BV of 100%). In Euryarchaeota, *Natronomonas pharaonis*, *Halorubrum lacusprofundi* and *Haloquadratum walsbyi* are grouped with other Halobacteriales (BV of 100%). The four Methanomicrobiales (*Methanocorpusculum labreanum*, *Methanospirillum hungatei*, *Candidatus* Methanoregula boonei and *Methanoculleus marisnigri*) are grouped together (BV of 100%) within a cluster that also contains Methanosarcinales (including their new representative *Methanosaeta thermophila*; BV of 100%) and Halobacteriales (BV of 100%). Finally, *Methanosphaera stadtmanae* emerges as a sister group of the other Methanobacteriale *Methanothermobacter thermautotrophicus* (BV of 100%), whereas *Methanococcus aeolicus* and *Methanococcus vannielii* are grouped with the other Methanococcales (BV of 100%).

In contrast to the tree that is based on SSU and LSU rRNA (FIG. 1), most relationships among the archaeal orders are well resolved and in agreement with previous studies[10], which highlights that R proteins are the phylogenetic markers of choice to study the archaeal phylogeny. Importantly, the monophylies of both hyperthermophilic crenarchaeota and Euryarchaeota are robustly recovered (each has a BV of 100%; FIG. 2). Interestingly, *C. symbiosum* constitutes a deeply branching lineage (BV of 99%), as it is a sister group of a clade that contains both Euryarchaeota (including *N. equitans*) and hyperthermophilic crenarchaeota. We think that this position is genuine and not the consequence of a long-branch attraction artefact, as the branch that leads to *C. symbiosum* is not particularly long

Clade
A monophyletic group.

Long-branch attraction artefact
A phylogenetic artefact that is induced by differences in evolutionary rates, and results in the artificial grouping of lineages that have long branches in a phylogenetic tree.

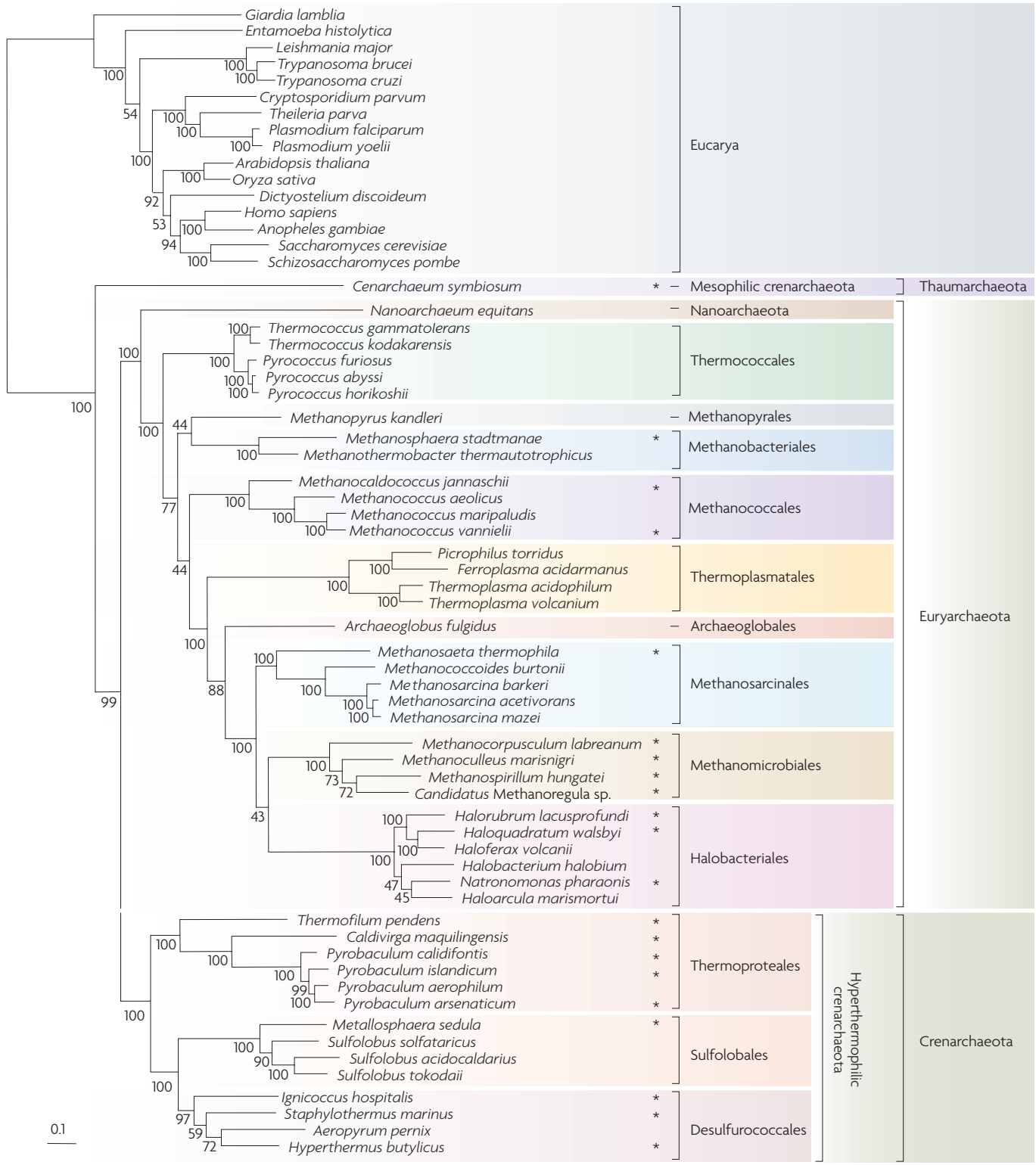


Figure 2 | **Maximum likelihood tree based on the concatenation of 53 R proteins from complete archaeal genomes.** Homologues of each R protein in complete genomes were retrieved by BLASTP and TBLASTN[60]. The concatenation included 53 alignments that harboured sequences from at least 61 of 64 taxa. The maximum likelihood phylogenetic tree was reconstructed using PHYML[61], with the Jones Taylor Thornton model of sequence evolution, by including a Γ-correction (eight categories of evolutionary rates, an estimated α-parameter and an estimated proportion of invariant sites). Numbers at nodes represent non-parametric bootstrap values computed by PHYML[61] (100 replications of the original dataset) using the same parameters. The use of different evolutionary models and methods did not produce differences in the resulting tree topology, at least for the archaeal part of the tree (not shown). Asterisks indicate the 21 new species (1 representative of the mesophilic crenarchaeota, *Cenarchaeum symbiosum*, 9 representatives of hyperthermophilic crenarchaeota and 11 representatives of Euryarchaeota) that were included in this analysis compared with previous work[11]. The scale bar represents the average number of substitutions per site.

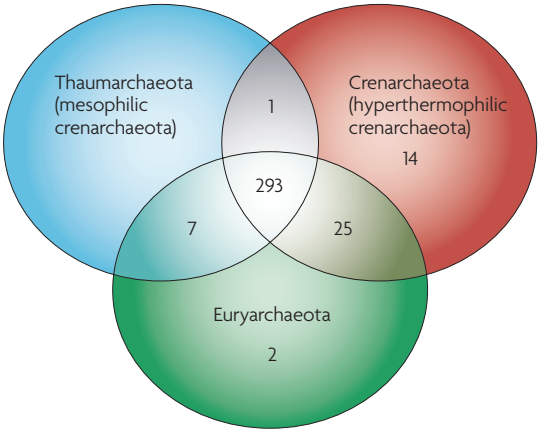


Figure 3 | **Scheme showing the number of proteins shared by Euryarchaeota, mesophilic crenarchaeota and hyperthermophilic crenarchaeota.**

in the tree (Fig. 2) or in individual R-protein trees (not shown), which indicates that its R proteins are not particularly fast-evolving. Moreover, even the fast-evolving Thermoplasmatales and lonely taxon *N. equitans* are not artificially attracted at the base of the tree (FIG. 2). However, a definitive exclusion of a long-branch attraction artefact[52] that could affect the position of *C. symbiosum* in this tree will only be possible by the addition of sequences from its relatives.

In conclusion, in contrast to SSU/LSU rRNA, analysis of R proteins improves the resolution of the deepest nodes in the archaeal phylogeny and suggests that mesophilic crenarchaeota could have diverged before the speciation of Euryarchaeota and hyperthermophilic crenarchaeota.

A conserved crenarchaeal genomic core?

Our SSU/LSU rRNA analysis only weakly suggests that mesophilic and hyperthermophilic crenarchaeota are sister groups (FIG. 1). By contrast, the analysis of R proteins indicates a robust and deeper branching of *C. symbiosum* that occurred before the speciation between Euryarchaeota and hyperthermophilic crenarchaeota (FIG. 2). This placement implies that mesophilic crenarchaeota are not more related to hyperthermophilic crenarchaeota than they are to Euryarchaeota. Thus, we investigated the presence in *C. symbiosum* of genes that seem to be strictly specific to Euryarchaeota (genes that are present in at least one representative of each major order of Euryarchaeota, but are absent from all representatives of Crenarchaeota); strictly specific to hyperthermophilic crenarchaeota (genes that are present in at least one representative of each major order of thermophilic crenarchaeota, but are absent from all representatives of Euryarchaeota); or that are common to Euryarchaeota and thermophilic crenarchaeota (FIG. 3). This criterion might seem stringent, as it excludes the markers that have been secondarily lost from some lineages (for example, histones in Thermoplasmatales). However, it has the advantage of focusing on genes that comprise the strictly conserved genomic core of Euryarchaeota and hyperthermophilic crenarchaeota, but avoiding the introduction of ambiguities that are due to genes with scattered distributions.

Using the NCBI COGs database (see Further information)[53], we identified 12 proteins that are strictly specific to Euryarchaeota, 15 proteins that are strictly specific to hyperthermophilic crenarchaeota (Supplementary information S2 (table)) and 318 proteins that are common to both phyla. Surprisingly, we found that *C. symbiosum* harbours 10 of the 12 euryarchaeal-specific proteins. Because HGTs from Euryarchaeota to mesophilic crenarchaeota were detected in a genome fragment from an uncultivated mesophilic crenarchaeon[32], we carried out a phylogenetic analysis of the ten euryarchaeal-specific proteins that were harboured by *C. symbiosum*. These trees, although generally poorly resolved (not shown), revealed that only three of these proteins might be present owing to HGT, whereas the remaining seven are probably ancestral traits that are common to Euryarchaeota and *C. symbiosum* (FIG. 3; Supplementary information S2 (table)). By contrast, *C. symbiosum* lacks 14 of the 15 hyperthermophilic crenarchaeal-specific proteins (including two R proteins) (FIG. 3; Supplementary information S2 (table)). Thus, with respect to the conserved genomic core, the mesophilic crenarchaeon *C. symbiosum* seems to be more similar to Euryarchaeota than to hyperthermophilic crenarchaeota. Importantly, a few of the euryarchaeal-specific genes that are present in *C. symbiosum* encode proteins that are involved in core cellular processes, such as DNA replication and cell division (Supplementary information S2 (table)), which shows that biologically important differences distinguish this organism, and by extension all mesophilic crenarchaeota, from hyperthermophilic crenarchaeota.

In addition to the presence of most euryarchaeal-specific proteins and absence of most proteins that are specific to hyperthermophilic crenarchaeota, *C. symbiosum* also lacks 25 proteins that are present in both Euryarchaeota and hyperthermophilic crenarchaeota, including the R protein S24e and the type I DNA topoisomerase of the A family (IA) (FIG. 3; Supplementary information S2 (table)). The absence of topoisomerase IA from *C. symbiosum* is surprising, as a protein from this family is present in representatives from the three domains of life[54], including archaea. Finally, *C. symbiosum* lacks the R protein L14e, which is present in all available genomes from hyperthermophilic crenarchaeota and basal euryarchaeota (Methanopyrales, Methanococcales, Methanobacteriales, Thermococcales and *N. equitans*), and the R protein L20a, which is present in all archaeal genomes except Thermoplasmatales. Moreover, we have identified potentially informative insertions and deletions (indels) in two other proteins, the R protein S27ae (hyperthermophilic crenarchaeota harbour a three-amino acid insertion that is absent from Euryarchaeota and mesophilic crenarchaeota) and the elongation factor EF-1α (both hyperthermophilic and mesophilic crenarchaeota harbour a conserved seven-amino acid insertion that is absent from Euryarchaeota). The distribution patterns of the features in the genome of *C. symbiosum* discussed above are puzzling, because they suggest that mesophilic crenarchaeota have a combination of traits that are either specific to hyperthermophilic crenarchaeota or Euryarchaeota.

Similar genome-mining data were recently obtained independently by Makarova, Koonin and co-workers[55], using an updated version of the NCBI COGs database that focused on Archaea. These authors noticed that the genome of *C. symbiosum* includes a much lower proportion of archaeal COGs than other archaeal genomes and groups with Euryarchaeota in a gene-content tree. They concluded from their analysis that "*C. symbiosum* is not a typical crenarchaeon (REF. 55)".

A third archaeal phylum?

Our SSU/LSU rRNA tree (FIG. 1) and analysis of the conserved genomic cores strongly reject the hypothesis that mesophilic crenarchaeota evolved from hyperthermophilic crenarchaeota (BV of 100%, which supports the monophyly of hyperthermophilic crenarchaeota). Moreover, our R-protein concatenation tree (FIG. 2) strongly rejects a sister-group relationship between hyperthermophilic crenarchaeota and *C. symbiosum*. Rather, it favours a deeper branching before the speciation of hyperthermophilic crenarchaeota and Euryarchaeota. The analysis of the genomic cores shows that *C. symbiosum* shares more features with Euryarchaeota than with hyperthermophilic crenarchaeota. This might indicate that *C. symbiosum* and its uncultivated relatives either belong to, or are sister to, Euryarchaeota. However, this is excluded by our phylogenetic analyses. Consistent with the basal emergence of mesophilic crenarchaeota, the genes of the euryarchaeal core that are shared with *C. symbiosum* can be interpreted as being ancestral characters that were present in the ancestor of archaea and were secondarily lost in the branch that led to hyperthermophilic crenarchaeota. We predict that the genomes of other mesophilic crenarchaeota from marine and terrestrial environments[56], such as *Candidatus* N. maritimus, will confirm our results when they become available for analysis. Moreover, this will enable the identification of features that are specific to the group, such as a conserved genomic core. One such feature could be the presence of a type I DNA topoisomerase of the B family (IB), which we detected in the genome of *C. symbiosum*. Whereas members of the topoisomerase IB family have never been identified in archaea, they are almost universal in eukarya and rarely present in bacteria[54]. This probably correlates with the absence from *C. symbiosum* of topoisomerase IA, which is present in all other archaea. Interestingly, the topoisomerase IB of *C. symbiosum* branches as a sister group to eukaryotes (not shown), which suggests that it was not transferred from the sponge host. A topoisomerase IB that was present in the last common ancestor of archaea and eukaryotes could later have been lost in the lineage that led to Euryarchaeota and hyperthermophilic crenarchaeota after their divergence from mesophilic crenarchaeota.

The diversity of mesophilic crenarchaeota based on SSU rRNA sequences[25,26,56,57] is comparable to that of hyperthermophilic crenarchaeota and Euryarchaeota, which suggests that they represent a major lineage that has equal status to Euryarchaeota and hyperthermophilic crenarchaeota. Indeed, environmental SSU rRNA surveys have already revealed several likely order-level subgroups within mesophilic crenarchaeota[25,26,56]. Moreover, the basal placement of one of their representatives in the archaeal phylogeny (FIG. 2) suggests that mesophilic crenarchaeota are an ancient lineage. This leads us to propose that mesophilic crenarchaeota represent a third archaeal phylum that we suggest naming the Thaumarchaeota (from the Greek '*thaumas*', meaning wonder). This choice was made to avoid any name that referred to phenotypic properties, such as mesophily, that could be challenged by the future identification of non-mesophilic organisms that belong to this phylum or the discovery of mesophilic relatives of cultivated hyperthermophilic crenarchaeota.

We stress that the classification of archaeal group I and its relatives as crenarchaeota was dubious from the outset, because their sequences formed only a sister group of hyperthermophilic crenarchaeota in the first rRNA trees[23]. The acceptance of this classification was probably influenced by the fact that the proposal to split the archaeal domain between Crenarchaeota and Euryarchaeota had only recently been made[6]. Clearly, the current classification of mesophilic crenarchaeota as Crenarchaeota is misleading, just as it is misleading to call methanogens 'methanogenic bacteria' because all methanogens are archaea. The proposal to establish mesophilic crenarchaeota as a third archaeal phylum goes beyond purely taxonomic purposes, and will stimulate research on this group of organisms and, more generally, on the Archaea.

Further phylogenetic analyses that include new members of the Thaumarchaeota are required to confirm the position of this phylum in the archaeal phylogeny. In any case, even if the basal branching of mesophilic crenarchaeota is challenged in favour of a sister grouping with hyperthermophilic crenarchaeota, this should not, in our opinion, change their phylum status, as they would remain a highly diversified and ancient group that have peculiar genomic characteristics. If the emergence of Thaumarchaeota prior to the speciation of Crenarchaeota and Euryarchaeota (as supported by R-protein analysis) is confirmed, this will leave open the nature of the last archaeal ancestor, which might have been either a mesophilic or psychrophilic organism (such as Thaumarchaeota) or a hyperthermophilic or thermophilic organism (such as cultivated crenarchaeota and some euryarchaeota). Importantly, the nature of the archaeal ancestor provides a different meaning for the HGTs from mesophilic euryarchaeota and bacteria to Thaumarchaeota that were highlighted from environmental genomics studies[32]. If the ancestor of Archaea was a hyperthermophile, HGT might have enabled the adaptation of hyperthermophilic thaumarchaeal lineages towards mesophily, as has been previously suggested[32]. Conversely, if the archaeal ancestor was a mesophile, HGT might have occurred between organisms that were thriving in the same low-temperature environments. Further studies on Thaumarchaeota will be essential to gain fundamental insights into the origin and early evolution of Archaea.

Mesophile
This term is normally restricted to organisms that have optimal growth temperatures of between 20 and 50°C. Here, however, the term mesophilic crenarchaeota is given to all non-hyperthermophilic crenarchaeota, even though some of them (presently uncultivated) are psychrophiles (optimal growth temperature of between 0 and 20°C) or moderate thermophiles (optimal growth temperature of between 50 and 70°C).

1. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
2. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090 (1977).
3. Fox, G. E. *et al.* The phylogeny of prokaryotes. *Science* **209**, 457–463 (1980).
4. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
5. Woese, C. R. in *Archaea: Evolution, Physiology and Molecular Biology* (eds Garrett, R. A. & Klenk, H. P.) 1–15 (Blackwell publishing, Oxford, 2006).
An eloquent historical review that reported, for the first time, all the steps that led to the discovery of Archaea. Forms the introduction of an interesting book that covers the various aspects of archaeal physiology and molecular biology and focuses on the similarities between Archaea and Eukaryotes.
6. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* **87**, 4576–4579 (1990).
7. Woese, C. R., Gupta, R., Hahn, C. M., Zillig, W. & Tu, J. The phylogenetic relationships of three sulfur dependent archaebacteria. *Syst. Appl. Microbiol.* **5**, 97–105 (1984).
8. Prangishvilli, D., Zillig, W., Gierl, A., Biesert, L. & Holz, I. DNA-dependent RNA polymerase of thermoacidophilic archaebacteria. *Eur. J. Biochem.* **122**, 471–477 (1982).
9. Makarova, K. S. & Koonin, E. V. Comparative genomics of archaea: how much have we learned in six years, and what's next? *Genome Biol.* **4**, 115 (2003).
10. Brochier, C., Forterre, P. & Gribaldo, S. An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol. Biol.* **5**, 36 (2005).
Revealed a conserved core of vertically inherited genes in different cellular systems, which proved that reconstructing the phylogeny of species is a feasible task in Archaea.
11. Gribaldo, S. & Brochier-Armanet, C. The origin and evolution of Archaea: a state of the art. *Phil. Trans. R. Soc. Lond. B* **361**, 1007–1022 (2006).
12. Daubin, V., Gouy, M. & Perriere, G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* **12**, 1080–1090 (2002).
13. Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1**, 8 (2001).
14. Forterre, P., Gribaldo, S. & Brochier-Armanet, C. in *Archaea: Evolution, Physiology and Molecular Biology* (eds Garrett, R. A. & Klenk, H. P.) 17–29 (Blackwell publishing, Oxford, 2006).
15. Bernander, R. Chromosome replication, nucleoid segregation and cell division in Archaea. *Trends Microbiol.* **8**, 278–283 (2000).
16. Myllykallio, H. *et al.* Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. *Science* **288**, 2212–2215 (2000).
17. Makarova, K. S. & Koonin, E. V. Evolutionary and functional genomics of the Archaea. *Curr. Opin. Microbiol.* **8**, 586–594 (2005).
18. Klenk, H. P. in *Archaea: Evolution, Physiology and Molecular Biology* (eds Garrity, G. M. & Klenk, H. P.) 75–95 (Blackwell publishing, Oxford, 2006).
19. Uemori, T., Sato, Y., Kato, I., Doi, H. & Ishino, Y. A novel DNA polymerase in the hyperthermophilic archaeon, *Pyrococcus furiosus*: gene cloning, expression, and characterization. *Genes Cells* **2**, 499–512 (1997).
20. Margolin, W., Wang, R. & Kumar, M. Isolation of an *ftsZ* homolog from the archaebacterium *Halobacterium salinarium*: implications for the evolution of FtsZ and tubulin. *J. Bacteriol.* **178**, 1320–1327 (1996).
21. Allers, T. & Mevarech, M. Archaeal genetics — the third way. *Nature Rev. Genet.* **6**, 58–73 (2005).
22. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.* **40**, 337–365 (1986).
23. DeLong, E. F. Archaea in coastal marine environments. *Proc. Natl Acad. Sci. USA* **89**, 5685–5689 (1992).
24. Fuhrman, J. A., McCallum, K. & Davis, A. A. Novel major archaebacterial group from marine plankton. *Nature* **356**, 148–149 (1992).
25. Schleper, C., Jurgens, G. & Jonuscheit, M. Genomic studies of uncultivated archaea. *Nature Rev. Microbiol.* **3**, 479–488 (2005).
A fascinating review that underlines the diversity of Archaea and the importance of using metagenomic approaches to gain insights into their biology and physiology.
26. Schleper, C. in *Archaea: Evolution, Physiology and Molecular Biology* (eds Garrett, R. A. & Klenk, H. P.) 39–50 (Blackwell publishing, Oxford, 2006).
27. Preston, C. M., Wu, K. Y., Molinski, T. F. & DeLong, E. F. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov. *Proc. Natl Acad. Sci. USA* **93**, 6241–6246 (1996).
28. Schleper, C., Swanson, R. V., Mathur, E. J. & DeLong, E. F. Characterization of a DNA polymerase from the uncultivated psychrophilic archaeon *Cenarchaeum symbiosum*. *J. Bacteriol.* **179**, 7803–7811 (1997).
29. Garrett, R. A. & Klenk, H. P. (eds) *Archaea: Evolution, Physiology and Molecular Biology* (Blackwell publishing, Oxford, 2006).
30. Barns, S. M., Delwiche, C. F., Palmer, J. D. & Pace, N. R. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl Acad. Sci. USA* **93**, 9188–9193 (1996).
31. Hershberger, K. L., Barns, S. M., Reysenbach, A. L., Dawson, S. C. & Pace, N. R. Wide diversity of Crenarchaeota. *Nature* **384**, 420 (1996).
32. Lopez-Garcia, P., Brochier, C., Moreira, D. & Rodriguez-Valera, F. Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers. *Environ. Microbiol.* **6**, 19–34 (2004).
33. Robertson, C. E., Harris, J. K., Spear, J. R. & Pace, N. R. Phylogenetic diversity and ecology of environmental Archaea. *Curr. Opin. Microbiol.* **8**, 638–642 (2005).
A recent, exhaustive archaeal phylogeny that was based on an analysis of SSU rRNA sequences. Despite the use of a large number of sequences, this marker was unable to resolve the deepest nodes of the archaeal phylogeny.
34. Cubonova, L., Sandman, K., Hallam, S. J., Delong, E. F. & Reeve, J. N. Histones in *Crenarchaea*. *J. Bacteriol.* **187**, 5482–5485 (2005).
35. Ochsenreiter, T., Selezi, D., Quaiser, A., Bonch-Osmolovskaya, L. & Schleper, C. Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR. *Environ. Microbiol.* **5**, 787–797 (2003).
36. Wuchter, C. *et al.* Archaeal nitrification in the ocean. *Proc. Natl Acad. Sci. USA* **103**, 12317–12322 (2006).
37. Leininger, S. *et al.* Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442**, 806–809 (2006).
38. Konneke, M. *et al.* Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**, 543–546 (2005).
The first report of the isolation of a member of the Thaumarchaeota, and the first demonstration of the ability of an isolated archaeon to oxidize ammonium.
39. Nunoura, T. *et al.* Genetic and functional properties of uncultivated thermophilic crenarchaeotes from a subsurface gold mine as revealed by analysis of genome fragments. *Environ. Microbiol.* **7**, 1967–1984 (2005).
40. Woese, C. R., Achenbach, L., Rouviere, P. & Mandelco, L. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. *Syst. Appl. Microbiol.* **14**, 364–371 (1991).
41. Boussau, B. & Gouy, M. Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* **55**, 756–768 (2006).
42. Hallam, S. J. *et al.* Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc. Natl Acad. Sci. USA* **103**, 18296–18301 (2006).
43. Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.* **19**, 631–639 (2002).
44. Waters, E. *et al.* The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. *Proc. Natl Acad. Sci. USA* **100**, 12984–12988 (2003).
45. Slesarev, A. I. *et al.* The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. *Proc. Natl Acad. Sci. USA* **99**, 4644–4649 (2002).
46. Gribaldo, S. & Philippe, H. Ancient phylogenetic relationships. *Theor. Popul. Biol.* **61**, 391–408 (2002).
47. Huber, H. *et al.* A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).
48. Brochier, C., Gribaldo, S., Zivanovic, Y., Confalonieri, F. & Forterre, P. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol.* **6**, R42 (2005).
49. Burggraf, S., Stetter, K. O., Rouviere, P. & Woese, C. R. *Methanopyrus kandleri*: an archaeal methanogen unrelated to all other known methanogens. *Syst. Appl. Microbiol.* **14**, 346–351 (1991).
50. Brochier, C., Forterre, P. & Gribaldo, S. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. *Genome Biol.* **5**, R17 (2004).
51. Bapteste, E., Brochier, C. & Boucher, Y. Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea* **1**, 353–363 (2005).
52. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410 (1978).
53. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28 (2001).
54. Forterre, P., Gribaldo, S., Gadelle, D. & Serre, M. C. Origin and evolution of DNA topoisomerases. *Biochimie* **89**, 427–446 (2007).
55. Makarova, K. S., Wolf, Y. I., Sorokin, A. V. & Koonin, E. V. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct* **2**, 33 (2007).
A dedicated COG database for Archaea that highlights the important differences between Thaumarchaeota and hyperthermophilic crenarchaeota.
56. Nicol, G. W. & Schleper, C. Ammonia-oxidising Crenarchaeota: important players in the nitrogen cycle? *Trends Microbiol.* **14**, 207–212 (2006).
Together with reference 37, this review revealed the unsuspected role of archaea in the global nitrogen cycle, which was previously assumed to be carried out by bacteria.
57. Forterre, P., Brochier, C. & Philippe, H. Evolution of the Archaea. *Theor. Popul. Biol.* **6**, 409–422 (2002).
58. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
59. Philippe, H. MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res.* **21**, 5264–5272 (1993).
60. Altschul, S. F. & Koonin, E. V. Iterated profile searches with PSI-BLAST — a tool for discovery in protein databases. *Trends Biochem. Sci.* **23**, 444–447 (1998).
61. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).

Acknowledgments
The authors thank G. Sezonov for suggesting the name Thaumarchaeota, E. Koonin for unpublished communications and the referees for useful comments and suggestions.

DATABASES

Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Aeropyrum pernix | *Caldivirga maquilingensis* | *Candidatus Methanoregula boonei* | *Cenarchaeum symbiosum* | *Halorubrum lacusprofundi* | *Haloquadratum walsbyi* | *Hyperthermus butylicus* | *Ignicoccus hospitalis* | *Metallosphaera sedula* | *Methanococcus aeolicus* | *Methanococcus vannielii* | *Methanocorpusculum labreanum* | *Methanoculleus marisnigri* | *Methanopyrus kandleri* | *Methanosaeta thermophila* | *Methanosphaera stadtmanae* | *Methanospirillum hungatei* | *Methanothermobacter thermautotrophicus* | *Nanoarchaeum equitans* | *Natronomonas pharaonis* | *Pyrobaculum aerophilum* | *Pyrobaculum arsenaticum* | *Pyrobaculum calidifontis* | *Pyrobaculum islandicum* | *Staphylothermus marinus* | *Thermofilum pendens*

FURTHER INFORMATION

Céline Brochier-Armanet's homepage: http://www.frangun.org
Céline Brochier-Armanet's laboratory website: http://lcb.cnrs-mrs.fr/
NCBI COGs database: http://www.ncbi.nlm.nih.gov/COG/

SUPPLEMENTARY INFORMATION

See online article: S1 (table) | S2 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

6

Pattern, Process, and the Early Evolution of Temperature on Earth

If nhPhyML could benefit from improvements in its capacity to explore the space of tree topologies, it is however able to reconstruct part of the process of evolution, and notably how the sequence G+C content evolved. I used this program to reconstruct the evolution or rRNA G+C contents along the tree of life. In parallel, Samuel Blanquart and Nicolas Lartillot analysed the evolution of protein sequence composition. Correlations between sequence composition and growth temperature found by Anamaria Necşsulea allowed us to propose a scenario for the evolution of growth temperatures along the tree of life.

Our results led us to consider the geological record in search of traces of events that might have caused the evolutions that we observed. Other studies might benefit from the associations of these two disciplines.

This article has been accepted for publication in *Nature*.

Accompanying Supplementary Materials can be found at the following address:

`http://biomserv.univ-lyon1.fr/~boussau/Article3/SupplementaryMaterial.pdf`

doi:10.1038/nature07393

LETTERS

Parallel adaptations to high temperatures in the Archaean eon

Bastien Boussau[1]*, Samuel Blanquart[2]*, Anamaria Necsulea[1], Nicolas Lartillot[2]† & Manolo Gouy[1]

Fossils of organisms dating from the origin and diversification of cellular life are scant and difficult to interpret[1], for this reason alternative means to investigate the ecology of the last universal common ancestor (LUCA) and of the ancestors of the three domains of life are of great scientific value. It was recently recognized that the effects of temperature on ancestral organisms left 'genetic footprints' that could be uncovered in extant genomes[2–4]. Accordingly, analyses of resurrected proteins predicted that the bacterial ancestor was thermophilic and that Bacteria subsequently adapted to lower temperatures[3,4]. As the archaeal ancestor is also thought to have been thermophilic[5], the LUCA was parsimoniously inferred as thermophilic too. However, an analysis of ribosomal RNAs supported the hypothesis of a non-hyperthermophilic LUCA[2]. Here we show that both rRNA and protein sequences analysed with advanced, realistic models of molecular evolution[6,7] provide independent support for two environmental-temperature-related phases during the evolutionary history of the tree of life. In the first period, thermotolerance increased from a mesophilic LUCA to thermophilic ancestors of Bacteria and of Archaea–Eukaryota; in the second period, it decreased. Therefore, the two lineages descending from the LUCA and leading to the ancestors of Bacteria and Archaea–Eukaryota convergently adapted to high temperatures, possibly in response to a climate change of the early Earth[1,8,9], and/or aided by the transition from an RNA genome in the LUCA to organisms with more thermostable DNA genomes[10,11]. This analysis unifies apparently contradictory results[2–4] into a coherent depiction of the evolution of an ecological trait over the entire tree of life.

Investigations into whether the LUCA was a hyperthermophilic (optimal growth temperature (OGT) $\geq$80 °C), thermophilic (OGT 50–80 °C), or mesophilic (OGT $\leq$50 °C) organism have relied on correlations between the species' OGT and the composition of their macromolecular sequences. In extant prokaryotic species, the G+C content of rRNA stems (that is, double-stranded parts) has been shown to correlate with OGT[12]. Exploiting this correlation, support was obtained for a non-hyperthermophilic LUCA[2]. In contrast, studies based on correlations between the composition of the LUCA's proteins and OGT concluded in favour of a hyperthermophilic LUCA[13,14] and of hyperthermophilic ancestors for both Archaea and Bacteria. The discrepancy between these results could come from some unexplained incongruence between rRNA and proteins, or, as we shall see, from differences between evolutionary models used.

These previous investigations[2,13,14] based their conclusions on comparisons of reconstructed ancestral sequence compositions with extant ones. Accurate modelling of the evolution of compositions is therefore crucial for such approaches. Two of these studies[13,14] relied on homogeneous models of evolution which make the simplifying hypothesis that substitutions occur with constant probabilities over time and across all lineages. If genomes and proteins had evolved according to a homogeneous model, they would all share the same base and amino acid compositions. Clearly, rRNA[12] and protein sequences[15] do not. Another approach[2] has been to use a branch-heterogeneous model of RNA sequence evolution. Branch-heterogeneous models are computationally more challenging, but more realistic as they allow replacement or substitution probabilities to vary between lineages, and thus explicitly account for compositional drifts[2,6,7,16,17]. Accordingly, they have been shown to accurately reconstruct ancestral sequence compositions[7].

We recently developed nhPhyML[7], an efficient program for the branch-heterogeneous modelling of nucleotide sequence evolution in the maximum likelihood framework, and nhPhyloBayes[6], which implements a site- and branch-heterogeneous Bayesian model of protein sequence evolution. The latter combines the break-point approach[17] to model variations of amino acid replacement rates along branches and the CAT[18] mixture model to account for site-wise variations of these rates. These models have been shown to describe the evolution of real sequences more faithfully than homogeneous ones[6,17], although neither homogeneous nor heterogeneous models ensure that inferred ancestral sequences are biologically functional. Using nhPhyML and nhPhyloBayes, we can reconstruct ancestral sequences of both rRNAs and proteins with branch-heterogeneous models, and estimate sequence compositions of all nodes of the tree of life, including the LUCA and its descendants. These compositions can be translated into approximate OGTs using the OGT/composition correlations observed in extant sequences[12,15].

A nucleotide data set of concatenated small- and large-subunit rRNAs—restricted to double-stranded regions—from 456 organisms (1,043 sites), and an amino acid data set of 56 concatenated nearly universal proteins from 30 organisms (3,336 sites), were assembled, each data set sampling all forms of cellular life. Correspondence analyses of the protein data set show that eukaryotes and prokaryotes markedly differ in amino acid compositions and that an effect of temperature on proteomes is detectable only among prokaryotic species (Supplementary Figs 4 and 6b). Similarly, the correlation between rRNA G+C content and OGT has only been documented in prokaryotes[12]. The ability to infer ancestral OGTs from rRNA and protein compositions therefore applies only to prokaryotes. However, eukaryotic sequences were kept in the subsequent analyses because they are part of the tree of life and as such provide useful phylogenetic information for ancestral sequence inferences.

The effect of temperature on prokaryotic proteomes is independent from genomic G+C contents[15], and was summarized in terms of average content in the amino acids I, V, Y, W, R, E and L (hereafter referred to as IVYWREL). Accordingly, our correspondence analysis identifies two independent factors accounting for most of the variance in amino acid compositions of prokaryotic proteins (Supplementary Fig. 5). The first factor (45.4% of the variance) highly correlates to

[1]Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université de Lyon, Université Lyon I, 43 Boulevard du 11 Novembre, 69622 Villeurbanne, France. [2]LIRMM, CNRS, 161 rue Ada, 34392 Montpellier, France. †Present address: Département de Biochimie, Université de Montréal, C.P. 6128, succursale Centre-Ville, Montréal QC H3C3J7, Canada.
*These authors contributed equally to this work.

genome G+C content ($r = 0.81$); the second (13.8% of the variance) is strongly correlated to OGT ($r = 0.83$) and to IVYWREL content ($r = 0.73$, Supplementary Fig. 6). The second factor was therefore used here as a molecular thermometer. The rRNA-based and the protein-based thermometers are thus independent, both because they come from distinct genome parts and because they exploit different effects of temperature on sequence composition. Furthermore, the correlation between rRNA G+C content and OGT is not expected to vary during evolutionary time because it stems from the different thermal stabilities of G–C and A–U RNA base pairs[12]. Thus, assuming that the relationship between temperature and amino acid composition of prokaryotes has also not varied since LUCA, the estimations of rRNA G+C content and amino acid compositions through branch-heterogeneous models provide two independent means to analyse the evolution of thermophily.

For each data set, a phylogenetic tree was inferred and rooted on the branch separating Bacteria from Archaea and Eukaryota (Supplementary Figs 7 and 8). Because the location of the root in the universal tree remains uncertain[19], the alternative rooting on the eukaryotic branch was also considered. Correlations between G+C content and OGT (Fig. 1a), and between the second axis of the amino acid correspondence analysis and OGT (Fig. 1b), were used to estimate OGTs for the LUCA and its descendants (Fig. 2).

Proteins and rRNAs support similar patterns of OGT changes for prokaryotes, so the discrepancy between previous rRNA- and protein-based investigations[2,13,14] was not a result of incongruence between these molecules. Protein-derived temperature estimates are generally lower than those based on rRNAs (Fig. 1), although some protein and rRNA-based OGT estimates overlap if confidence intervals of ancestral compositions are taken into account (Supplementary Table 3). Both types of data support key conclusions (Fig. 1). First, the LUCA is predicted to be a non-hyperthermophilic organism, as previously reported[2]. Second, both archaeal and bacterial ancestors, as well as the common ancestor of Archaea and Eukaryota, are estimated to have been thermophilic to hyperthermophilic (Fig. 2). This result is in line with previous studies[3,5]. Third, within the bacterial phylogenetic tree, tolerance to heat decreased (Fig. 2). This last result is congruent with recent estimates of the evolution of OGTs in the bacterial domain based on ancestral reconstructions and characterizations of elongation factor Tu proteins[4].

Support for the hypothesis of a non-hyperthermophilic LUCA and of subsequent parallel adaptations to high temperatures partly rests on a protein content depleted in IVYWREL for the LUCA and subsequently enriched in these amino acids. This is consistent with a recent report that amino acids IVYEW might be under-represented in LUCA's proteins[20]. This finding has been interpreted as evidence that these five amino acids were a late addition to the genetic code,

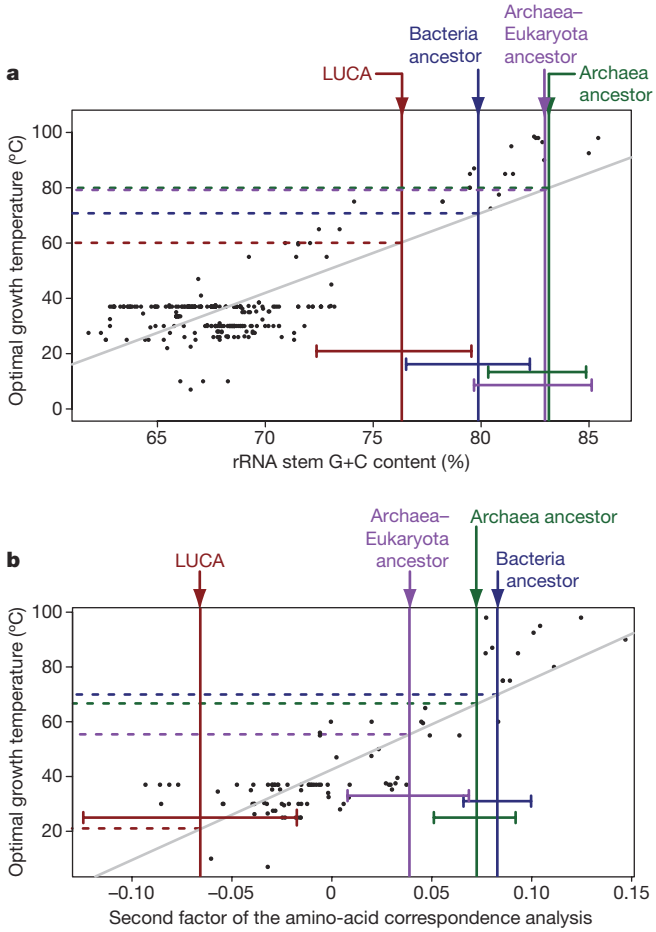


Figure 1 | Correlations between sequence compositions and OGT, and estimates of key ancestral compositions. Black dots indicate extant prokaryotes positioned according to their sequence composition and OGT. Dashed coloured lines indicate predicted OGTs for various ancestors. **a**, Correlation between rRNA G+C content and OGT. The vertical coloured bars indicate most likely nhPhyML estimates of ancestral G+C contents with their 95% confidence intervals. **b**, Correlation between the second factor of the correspondence analysis on amino acid compositions and OGT. The vertical coloured bars indicate median ancestral compositions inferred by nhPhyloBayes with their 95% confidence intervals. The LUCA is significantly less thermophilic than its direct descendants ($P \leq 0.005$).

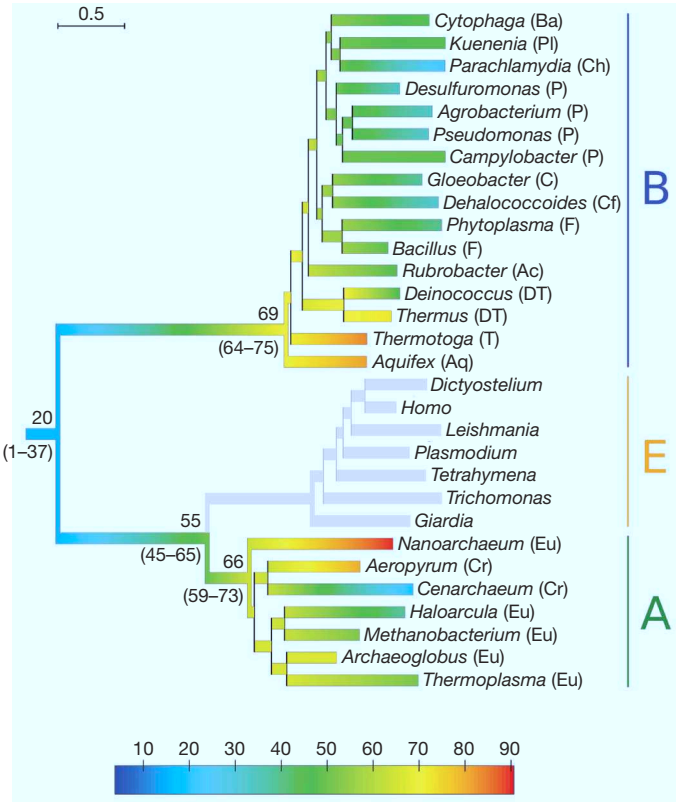


Figure 2 | Evolution of thermophily over the tree of life. Protein-derived nhPhyloBayes OGT estimates (and their 95% confidence intervals for key ancestors) for prokaryotic organisms are colour-coded from blue to red for low to high temperatures. Colours were interpolated between temperatures estimated at nodes. The eukaryotic domain, in which OGT cannot be estimated, has been shaded. The colour scale is in °C; the branch length scale is in substitutions per site. A, archaeal; B, bacterial; E, eukaryotic domains. Ac, Actinobacteria; Aq, Aquificae; Ba, Bacteroidetes; C, Cyanobacteria; Cf, Chloroflexi; Ch, Chlamydiae; Cr, Crenarchaeota; DT, Deinococcus/Thermus; Eu, Euryarchaeota; F, Firmicutes; P, Proteobacteria; Pl, Planctomycetes; T, Thermotogae.

and that the proteome of the LUCA had not yet reached compositional equilibrium. Although such interpretation in terms of early genetic code evolution is possible, our hypothesis of parallel adaptations to high temperatures has the advantage of explaining the patterns observed with both rRNAs and proteins.

Additional experiments suggest that the present analyses of rRNA and protein sequences with branch-heterogeneous models of evolution uncover genuine signals of ancient temperature preferences and are not affected by systematic biases.

First, these results are robust to changes in the topology chosen for inference because analyses with alternative topologies yielded virtually identical OGT estimates (Supplementary Fig. 10). Moreover, phylogenetic trees rooted on the eukaryotic branch also suggest that OGT increased between the universal ancestor and the divergence of Archaea and Bacteria (Supplementary Figs 13–15).

Second, taxonomic sampling does not strongly affect these results. With rRNA and protein data sets in which eukaryotic sequences were removed, the signal for OGT increases between the LUCA and the domain ancestors was essentially unchanged (Supplementary Fig. 36). Moreover, both for rRNAs and proteins, two artificially biased data sets containing sequences from either thermophilic or mesophilic prokaryotes were assembled (see Supplementary Information). The signal for parallel increases in OGT is confirmed in all but one of these four data sets: the mesophilic rRNA data set. However, the longest of the two mesophilic alignments, the protein data set, supports the same pattern of OGT changes as the complete data sets (Supplementary Figs 16 and 17). Notably, analysis of the protein mesophilic data set shows that this pattern is independent of the debated position of hyperthermophilic organisms in the tree of life. Furthermore, with all rRNA and protein data sets, even with the sampling limited to thermophilic prokaryotes, the LUCA remains predicted as a non-hyperthermophilic organism (Supplementary Figs 18 and 19).

Third, dependence of the results on models used for ancestral reconstruction was investigated. Additional branch-heterogeneous evolutionary models were applied, two to the rRNA data set, and one to the protein data set (see Supplementary Information). All these alternative branch-heterogeneous models confirm our results (Supplementary Figs 21–23, 29 and 30). Compositional analyses were also conducted using branch-homogeneous models of evolution: GTR[21] for rRNA and proteins, and CAT[18] for proteins. All these models tend to predict parallel adaptations to higher temperatures from the LUCA to its descendants, suggesting the existence of a genuine signal for such a pattern in the data (Supplementary Figs 24, 26 and 28). However, only when models are realistic enough is the LUCA predicted as significantly less thermophilic than its two descendants. For instance, ancestral protein compositions predicted by the GTR model for the LUCA and its two descendants strongly overlap, which may explain previously published results[13], whereas the CAT model better separates these ancestral node distributions, although less clearly than does the CAT–BP branch-heterogeneous model (Supplementary Figs 26, 28 and 29). These experiments show that as the evolutionary process is more accurately modelled, the support for parallel increases in OGT from the LUCA to its offspring is strengthened.

Fourth, it is known that the base compositions of fast and slowly evolving sites and, particularly, of single- and double-stranded regions of rRNA molecules differ and that this may bias ancestral sequence estimates[16]. To minimize this bias, only double-stranded rRNA regions have been analysed here. Moreover, if fast-evolving sites are removed, estimates still support parallel adaptations to high temperatures (Supplementary Fig. 33).

Fifth, it has been shown that some ancestral reconstruction methods might improperly estimate the frequencies of rare amino acids[22]. To control for that potential bias, the two rarest amino acids, cysteine and tryptophan, were discarded from estimated ancestral sequences: this had essentially no impact on results (Supplementary Fig. 34).

Sixth, the sensitivity of the OGT estimates at the tree root to the prior distribution of ancestral amino acid compositions used for Bayesian analyses was investigated (Supplementary Fig. 35). This prior distribution induces a flat, uninformative distribution over OGTs, whereas the posterior distributions estimated for LUCA and the bacterial ancestor have small variance, and thus reflect a genuine signal in the data, rather than a bias from the prior. Moreover, even with a strongly informative prior distribution that is biased towards high temperature amino acid distributions, the posterior distribution of the LUCA's amino acid composition, although altered, is centred at lower temperatures than that of the bacterial ancestor.

The present use of molecular thermometers requires that evolution of the data sets under analysis can be modelled by a tree structure as far as reconstruction of ancestral compositions is concerned. We emphasize that our protein analyses are based on 56 genes that did not undergo between-domain transfers (see Methods), which precludes that ancestral sequence reconstructions are confounded by such gene exchanges. We do not exclude within-domain lateral transfers of these genes; however, the robustness of the inferred ancestral compositions to alternative domain phylogenies[4,7] (see also Supplementary Figs 10 and 20) suggests that these potential transfers do not fundamentally affect the results for domain ancestors. Finally, because molecular thermometers measure the average environmental temperature of the hosts of ancestral genes, they apply even if ancestral genes of extant prokaryotes originate from diverse organisms[19].

Thus, all our analyses support the hypothesis of a non-hyperthermophilic LUCA and of transitions to higher environmental temperatures for its descendants. Although these organisms have not yet been anchored in time[23], a few geological and biological factors may explain observed changes in temperature preferences. It has already been observed[4] that the general trend of decreasing OGTs from the bacterial ancestor to extant species strikingly parallels recent geological estimates of the progressive cooling down of oceans shifting from about 70 °C 3.5 billion years ago to approximately 10 °C at present[24]. The evolution of thermophily in the bacterial domain might therefore stem from the continuous adjustment of Bacteria to ocean temperatures, although the evidence for a hot Archaean climate remains debated[25]. A similar conclusion may apply to Archaea as well, but would require confirmation with additional genome sequences from mesophilic Archaea. A hot Archaean ocean may preclude the existence of a cool 'little pond' where the LUCA could have evolved. Therefore, a non-hyperthermophilic LUCA would suggest that moderate temperatures existed earlier in the history of the Earth.

Geological data about palaeoclimates that old are very scarce. However, some models of Hadean and early Archaean climates (3.5–4.2 billion years ago) suggest that the Earth might have been colder than it is today, possibly covered with frozen oceans[1,26]. Moreover, a hypothesis of brutal temperature changes involving meteoritic impacts that boiled the oceans and therefore nearly annihilated all life forms but the most heat-resistant ones has been proposed[1,8,9]. Huge meteorites probably impacted the Earth at least as late as 3.8–4 billion years ago, most notably during the late heavy bombardment[27] and created a series of brief but very hot climates on Earth[1]. As life may have originated more than 3.7 billion years ago[28], it is possible that early organisms, namely the LUCA's offspring, experienced such bottlenecks.

Alternatively, under the hypothesis that life originated extra-terrestrially, the transfer of life to the Earth from another planet in ejecta created by meteorite impacts would have also entailed selection of heat-resistant cells[1]. Overall, geological knowledge provides several frames that might fit the predictions of our biological thermometers.

A biological hypothesis could provide an internal mechanism to explain the observed pattern. It posits that the LUCA had an RNA genome, and that its offspring lineages independently evolved the ability to use DNA for genome encoding[10], possibly by co-opting it from viruses[11]. Although our results do not bring direct evidence in support of this hypothesis, they are compatible with it and could even

help explain such independent acquisitions of DNA in adaptive terms, as DNA is much more thermostable than RNA[29].

Great care is necessary when attempting a reconstruction of events that took place more than three billion years ago. However, the strong agreement between results obtained using two types of data (proteins and rRNAs), two independent temperature proxies (protein amino acid composition and rRNA G+C content), and independently developed statistical models, is remarkable. This suggests that a similar approach could successfully be used to gain insight into other ecological features of early life. For example, it has been shown that aerobic and anaerobic bacteria differ in the amino acid composition of their proteome[30]; future ancestral sequence reconstructions could reveal the evolution of aerobiosis along the tree of life in relation with the geological record of oxygen atmospheric concentration.

METHODS SUMMARY

Ribosomal RNA sequences were aligned according to their shared secondary structure. Sites belonging to double-stranded stems were selected to obtain an alignment of 1,043 stem sites for 456 organisms. Protein families with wide species coverage and no or very low redundancy in all species were selected from the HOGENOM database of families of homologous genes. Only sites showing less than 5% gaps were kept, giving an alignment of 3,336 positions for 30 organisms. Phylogenetic trees were inferred using Bayesian or maximum likelihood techniques. Ancestral nucleotide and amino acid compositions were inferred for all tree nodes using the programs nhPhyML[7] and nhPhyloBayes[6], respectively. The G+C contents of ancestral rRNA sequences were compared to extant rRNA base compositions. The second factor of the correspondence analysis of amino acid compositions of extant prokaryotic proteins was used to estimate ancestral environmental temperatures by adding ancestral amino acid compositions as supplementary rows to the correspondence analysis. These two procedures allowed us to estimate ancestral environmental temperatures with the rRNA and the protein data sets, respectively. Confidence intervals for the estimated environmental temperatures were as follows: in the case of rRNAs, they contained 95% of the distribution obtained by a bootstrap procedure (200 replicates); for Bayesian analyses, regular 95% credibility intervals were computed from a sample of 2,000 points drawn from the posterior distribution.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 5 March; accepted 1 September 2008.
Published online 26 November 2008.

1. Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409,** 1083–1091 (2001).
2. Galtier, N., Tourasse, N. & Gouy, M. A nonhyperthermophilic common ancestor to extant life forms. *Science* **283,** 220–221 (1999).
3. Gaucher, E. A., Thomson, J. M., Burgan, M. F. & Benner, S. A. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425,** 285–288 (2003).
4. Gaucher, E. A. Govindara jan, S. & Ganesh, O. K. Palaeotemperature trend for precambrian life inferred from resurrected proteins. *Nature* **451,** 704–707 (2008).
5. Gribaldo, S. & Brochier-Armanet, C. The origin and evolution of archaea: a state of the art. *Phil. Trans. R. Soc. Lond. B* **361,** 1007–1022 (2006).
6. Blanquart, S. & Lartillot, N. A site- and time-heterogeneous model of amino-acid replacement. *Mol. Biol. Evol.* **25,** 842–858 (2008).
7. Boussau, B. & Gouy, M. Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* **55,** 756–768 (2006).
8. Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* **342,** 139–142 (1989).
9. Gogarten-Boekels, M., Hilario, E. & Gogarten, J. P. The effects of heavy meteorite bombardment on the early evolution–the emergence of the three domains of life. *Orig. Life Evol. Biosph.* **25,** 251–264 (1995).
10. Mushegian, A. R. & Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA* **93,** 10268–10273 (1996).
11. Forterre, P. The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.* **5,** 525–532 (2002).
12. Galtier, N. & Lobry, J. R. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J. Mol. Evol.* **44,** 632–636 (1997).
13. Di Giulio, M. The universal ancestor and the ancestor of bacteria were hyperthermophiles. *J. Mol. Evol.* **57,** 721–730 (2003).
14. Brooks, D. J., Fresco, J. R. & Singh, M. A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. *Bioinformatics* **20,** 2251–2257 (2004).
15. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput. Biol.* **3,** 62–72 (2007).
16. Gowri-Shankar, V. & Rattray, M. On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Mol. Biol. Evol.* **23,** 352–364 (2005).
17. Blanquart, S. & Lartillot, N. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.* **23,** 2058–2071 (2006).
18. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21,** 1095–1109 (2004).
19. Zhaxybayeva, O., Lapierre, P. & Gogarten, J. P. Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* **227,** 53–64 (2005).
20. Fournier, G. P. & Gogarten, J. P. Signature of a primitive genetic code in ancient protein lineages. *J. Mol. Evol.* **65,** 425–436 (2007).
21. Lanave, C., Preparata, G., Saccone, C. & Serio, G. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20,** 86–93 (1984).
22. Williams, P. D., Pollock, D. D., Blackburne, B. P. & Goldstein, R. A. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput. Biol.* **2,** 598–605 (2006).
23. Graur, D. & Martin, W. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* **20,** 80–86 (2004).
24. Robert, F. & Chaussidon, M. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature* **443,** 969–972 (2006).
25. Shields, G. A. & Kasting, J. F. Evidence for hot early oceans? *Nature* **447,** E1 (2007).
26. Kasting, J. F. & Ono, S. Palaeoclimates: the first two billion years. *Phil. Trans. R. Soc. Lond. B* **361,** 917–929 (2006).
27. Gomes, R., Levison, H. F., Tsiganis, K. & Morbidelli, A. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature* **435,** 466–469 (2005).
28. Rosing, M. T. ^{13}C-depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from West Greenland. *Science* **283,** 674–676 (1999).
29. Islas, S., Velasco, A. M., Becerra, A., Delaye, L. & Lazcano, A. Hyperthermophily and the origin and earliest evolution of life. *Int. Microbiol.* **6,** 87–94 (2003).
30. Naya, H., Romero, H., Zavala, A., Alvarez, B. & Musto, H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J. Mol. Evol.* **55,** 260–264 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by Action Concertée Incitative IMPBIO-MODELPHYLO and ANR PlasmoExplore. We thank C. Brochier-Armanet and A. Lazcano for help and suggestions, the LIRMM Bioinformatics platform ATGC and the computing facilities of IN2P3.

Author Contributions B.B. and S.B. contributed equally to this study, designing and conducting experiments. A.N. performed statistical analyses and retrieved optimal growth temperatures. N.L. and M.G. provided guidance throughout the study, and M.G. gave the original idea. All authors participated in manuscript writing.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.G. (mgouy@biomserv.univ-lyon1.fr).

METHODS

rRNA data set. Prokaryotic small (SSU) and large (LSU) subunit rRNAs were retrieved in January 2007 from complete genomes available at the National Center for Biotechnology Information (NCBI). SSU and LSU rRNA sequences from ongoing genome projects or from large genomic fragments of important or poorly represented groups (for example, Archaea or hyperthermophilic bacteria) were added in June 2007. Eukaryotic SSU and LSU rRNA sequences were provided by D. Moreira; 65 slowly evolving sequences were selected from this data set[31]. Sequences were aligned using MUSCLE[32]. Resulting alignments were concatenated and manually improved using the MUST package[33]. Regions of doubtful alignment were removed using the MUST package; 2,239 sites were kept. A distance phylogenetic tree was computed using dnadist (Jukes and Cantor model) and neighbour from the PHYLIP package[34]. The final data set contained 65 eukaryotic, 60 archaeal and 331 bacterial sequences representative of the molecular diversity in each domain. An additional data set of 60 sequences sampling the diversity of the full data set was used in Bayesian analyses. Secondary structure predictions were downloaded from the rRNA database[35]. Sites that were predicted as double-stranded stems in *Saccharomyces cerevisiae, Escherichia coli* and *Archaeoglobus fulgidus* were selected to give an alignment of 1,043 sites.

Protein data set. Nearly universal protein families with one member per genome were used to avoid ill-defined orthology. Protein families from the HOGENOM database of families of homologous genes (release 03, October 2005, S. Penel and L. Duret, personal communication; http://pbil.univ-lyon1.fr/databases/hogenom3.html) that displayed a wide species coverage with no or very low redundancy in all species were selected. Additional sequences from other genomes whose phylogenetic position was interesting were considered. These were downloaded from the Joint Genome Institute (*Desulfuromonas acetoxidans*), The Institute for Genomic Research (*Giardia lamblia, Tetrahymena thermophila, Trichomonas vaginalis*) or the NCBI (*Kuenenia stuttgartiensis*), and were searched for homologous genes using BLAST[36]; only the best hit was retrieved. The protein families were subsequently aligned using MUSCLE[32] and submitted to phylogenetic analysis using the NJ algorithm[37] with Poisson distances with Phylo_Win[38]. Proteins from mitochondrial or chloroplastic symbioses and families in which horizontal transfers between Bacteria and Archaea may have occurred were discarded, and so were aminoacyl-tRNA synthetases prone to transfers[39]. In the rare families with two sequences from the same species, the sequence showing the longest terminal branch or whose position was most at odds with the biological classification was discarded. This provided 56 protein families (Supplementary Table 2) for 115 species, which were concatenated using ScaFos[40]. From the 9,218 concatenated sites, 3,336 positions with less than 5% gaps were conserved. The whole data set was used to compute the correspondence analysis and correlations between amino-acid composition and optimal growth temperature. For Bayesian analyses, 30 species among 115 were selected sampling the diversity of cellular life (Supplementary Table 1).

Multivariate data analyses. Correspondence analysis[41] was performed on the amino-acid compositions of the protein data set, using the ade4 package[42] of the R environment for statistical computing.

Phylogenetic tree construction. An rRNA phylogenetic tree was built from the 456-sequence alignment with both stems and loops with PhyML_aLRT[43,44] with the GTR model, a gamma law with eight categories and an estimated proportion of invariant sites. The tree for the 60-sequence data set was obtained in the same manner. The phylogenetic trees for the three protein data sets (Supplementary Table 1) were obtained using MrBayes 3.1.1 (ref. 45), using the GTR substitution model and a gamma law with four categories for rates across sites. Chains were run for 1,000,000 generations and samples were collected each 100 generations, a burn-in of 1,000 samples was discarded. The majority rule consensus was computed from the 9,000 remaining samples.

Identification of fast-evolving rRNA sites. Posterior probabilities for gamma law rate categories were predicted for each site with PhyML_aLRT. Site evolutionary rates were obtained by averaging gamma law rate categories weighted by their posterior probabilities. Sites whose evolutionary rate was above the arbitrarily chosen threshold of 2.0 (Supplementary Fig. 2) were discarded, which left 940 sites.

Estimation of ancestral compositions. For the maximum likelihood approach, nhPhyML[7] was applied to the rRNA stem sites alignment and the phylogenetic tree described above, and used to estimate all evolutionary parameter values, except tree topology, which was fixed. Site-specific ancestral nucleotide compositions at tree root and at internal node j descendant of node i were computed by:

$$p_{\mathrm{root}}(x) = a(x)L_{\mathrm{low}}(x \text{ at root})/L;\ a(\mathrm{A}) = a(\mathrm{T}) = (1-\omega)/2;\ a(\mathrm{C}) = a(\mathrm{G}) = \omega/2$$

$$p_j(x) = (\textstyle\sum_y L_{\mathrm{upp}}(y \text{ at node } i)\ p_{y\to x}\ L_{\mathrm{low}}(x \text{ at node } j))/L$$

where x and y are in{A, C, G, T}, L is the total tree likelihood at this site, L_{low} and L_{upp} are site lower and upper conditional likelihoods, respectively[7], ω is the maximum likelihood estimate of root G+C content, and $p_{y\to x}$ is the probability of the y to x substitution on the i to j branch. For Bayesian analyses, nhPhyloBayes[6] was applied to trees described above. Ancestral sequence reconstruction started, for each site, by drawing a state x at the root: $x \sim \omega(x)L_{\mathrm{low}}(x$ at root), where ω was the Markov Chain Monte Carlo[45] (MCMC) estimate of root amino acid or nucleotide frequencies. Then, states x have been recursively drawn at each node j: $x \sim p_{y\to x}L_{\mathrm{low}}(x$ at $j)$, where y was the parental node state. Given a realization of the model, this permitted the reconstruction of ancestral sequences at all nodes. Posterior distributions were sampled by 2 (for proteins) or 4 (for rRNA) independent MCMC chains, each with 1,000 to 2,000 realizations. Posterior distributions of sequence compositions combined all realizations of all chains. Protein ancestral compositions were projected on the second axis of the correspondence analysis, and rRNA ancestral compositions were summed up as G+C contents.

Statistical tests. In bootstrap analyses, all parameters but topology and branch lengths were estimated under the maximum likelihood criterion for each replicate. In tests of whether the LUCA is less thermophilic than one of its descendants, P values were the fraction of cases where the temperature estimate for LUCA in a bootstrap replicate or in an iteration of an MCMC chain was above the estimate obtained for its descendant.

31. Moreira, D. *et al.* Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. *Mol. Phylogenet. Evol.* **44,** 255–266 (2007).
32. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).
33. Philippe, H. MUST, a computer package of management utilities for sequences and trees. *Nucleic Acids Res.* **21,** 5264–5272 (1993).
34. Felsenstein, J. *PHYLIP (Phylogeny Inference Package) version 3.6.* (Department of Genome Sciences, 2005).
35. Wuyts, J., Perrière, G. & Van De Peer, Y. The European ribosomal RNA database. *Nucleic Acids Res.* **32,** D101–D103 (2004).
36. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
37. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).
38. Galtier, N., Gouy, M. & Gautier, C. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* **12,** 543–548 (1996).
39. Wolf, Y. I., Aravind, L., Grishin, N. V. & Koonin, E. V. Evolution of aminoacyl-trna synthetases–analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9,** 689–710 (1999).
40. Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* **7** (Suppl 1), S2 (2007).
41. Hill, M. O. Correspondence analysis: a neglected multivariate method. *Appl. Statist.* **23,** 340–354 (1974).
42. Chessel, D., Dufour, A. B. & Thioulouse, J. The ade4 package -I- one-table methods. *R. News* **4,** 5–10 (2004).
43. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52,** 696–704 (2003).
44. Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55,** 539–552 (2006).
45. Huelsenbeck, J. P. & Ronquist, F. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).

Further tests

The present article argues in favour of a non-parsimonious pattern, with parallel evolutions to high temperatures from LUCA to its descendants. Although several tests have been performed to ensure that an artifact was not at the origin of this pattern, we chose to make one more test, based on simulations. The question we wanted to ask was: if a particular pattern comparable to what has been found in the real data is simulated, can nhPhyML recover it? All we needed was a program able to make simulations easily, with a non-homogeneous model of sequence evolution. Such a program is presented in the next article (article 7).

We based our simulations on the results found by nhPhyML on the dataset containing 456 rRNA sequences 1043 bases long. On this dataset, nhPhyML predicted a G+C content at the root around 73%, and equilibrium frequencies towards the Bacterial and Archaea-Eukaryota ancestors of around 97% and 87%.

We therefore chose to simulate datasets as follows:

- We extracted a subtree of 20 leaves from the rRNA tree obtained in the article, using the program bppphysamp, from the Bio++ library (Dutheil *et al.*, 2006), so that the sampling homogeneously covers all the diversity present in the tree. We then used this tree topology to simulate datasets of 20 sequences 1043 bases long. For each dataset to simulate:

- We randomly pick a G+C content ω at the root, $\omega \in [0.2; 0.8]$. Base frequencies are obtained as follows: $[A] = 0.4 \times (1-\omega); [C] = 0.4 \times (\omega); [G] = 0.6 \times (\omega); [T] = 0.6 \times (1-\omega)$

- On each branch of the tree, a different HKY model (Hasegawa *et al.*, 1985) is used. An equilibrium G+C content θ is defined, $\theta \in [0.1; 0.9]$, and equilibrium base frequencies are obtained as for the root base frequencies. Because we want to simulate datasets comparable to the real one, for the two branches coming from the root, we make sure that the absolute difference between θ and ω is superior or equal to 0.15.

- Sequences are evolved given the root base compositions and the models on each branch. During this evolution, a discretized gamma law (Yang, 1994) with 4 categories plus a category of invariant is used, with an alpha parameter set to 1.0, and the transition/transversion ratio is set to 1.0.

We ran 150 such simulations. We then used nhPhyML on each of these simulations (using a gamma law with four categories, and estimated alpha and transition/transversion parameters) with the true topology, and studied how nhPhyML reconstructed the evolution of G+C content in LUCA and its two descendants.

We defined three possible categories for the simulations, depending on the G+C content evolution from LUCA to its descendants:

- Parallel increases in G+C content (I)
- Parallel decreases in G+C content (D)
- One increase and one decrease in G+C content (ID)

We compared the results obtained by nhPhyML to what had been simulated, and obtained the following results:

Category	I	D	ID
Proportion of correct reconstruction	53/56 (95%)	53/58 (91%)	33/36 (92%)

The above table shows that nhPhyML accurately reconstructs the true evolutionary scenario. In all the rare cases where it fails to recognize parallel evolutions, it recovers a situation where there is an increase in a branch, and a decrease in the other branch (situation ID).

Interestingly, if we use the same simulation protocol, but this time do not use a gamma law in nhPhyML to reconstruct the evolutionary scenario, the program's accuracy considerably drops, as shown in the next table:

Category	I	D	ID
Proportion of correct reconstruction	51/56 (91%)	49/55 (89%)	26/39 (67%)

The drop is particularly striking when the root G+C content is between the G+C contents of its descendants (situation ID), perhaps because in this case, the signal for irreversibility is more feeble.

This reveals that the model uses rate heterogeneities to recover the true evolutionary history in deep branches: in such circumstances, the model can distinguish slowly evolving from fast evolving sites and infer frequencies at the root more reliably by trusting slow sites against fast ones.

This intuition is confirmed if we build consensus sequences of the rRNA alignment. When a position should be identical in 80% of the sequences to enter the consensus, the consensus sequence has a G+C content of $\approx 74.0\%$ (and a length of 584 sites); if the threshold is put at 90% instead, the consensus sequence has a G+C content of $\approx 71.4\%$ (for a length of 419 sites). The consensus obtained at the 90% threshold contains sites that have undergone less substitutions and are less G+C rich than the consensus obtained at the 80% threshold. Part of the

signal for a LUCA rRNA less G+C rich than its two descendants may therefore come from the fact that slowly evolving sites in rRNA are less G+C rich than faster sites.

The situation is probably the same for protein sequences, as Fournier et Gogarten (2007) found that the amino-acids occuring in higher frequencies in heat-loving organisms occur in lower proportions among constant positions than in other positions. Non-homogeneous models of evolution thus find signal for ancestral compositions where it lies, among fossil sites.

7

Towards Better Non-Homogeneous Models of Sequence Evolution

Previous works convinced me that more research needed to be done on models able to cope with composition heterogeneity. The Bio++ libraries are a set of C++ routines for making phylogenetic analyses, notably. Julien Dutheil and I implemented support for non-homogeneous models in Bio++. This should help researchers experiment with new models or new algorithms, and may democratise such models.

This article has been accepted in *BMC Evolutionary Biology*.

BMC Evolutionary Biology

BioMed Central

Software

Open Access

Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs

Julien Dutheil*[1] and Bastien Boussau[2]

Address: [1]BiRC – Bioinformatics Research Center – University of Aarhus, C. F. Møllers Alle, Building 1110, DK-8000 Århus C, Denmark and [2]UMR CNRS 5558 – Laboratoire de Biométrie et Biologie Évolutive, CNRS, Université de Lyon, Université lyon 1, 43 Boulevard du 11 Novembre, 69622 Villeurbanne, France

Email: Julien Dutheil* - jdutheil@daimi.au.dk; Bastien Boussau - boussau@biomserv.univ-lyon1.fr

* Corresponding author

Published: 22 September 2008

BMC Evolutionary Biology 2008, **8**:255 doi:10.1186/1471-2148-8-255

Received: 24 April 2008
Accepted: 22 September 2008

This article is available from: http://www.biomedcentral.com/1471-2148/8/255

Abstract

Background: Accurately modeling the sequence substitution process is required for the correct estimation of evolutionary parameters, be they phylogenetic relationships, substitution rates or ancestral states; it is also crucial to simulate realistic data sets. Such simulation procedures are needed to estimate the null-distribution of complex statistics, an approach referred to as parametric bootstrapping, and are also used to test the quality of phylogenetic reconstruction programs. It has often been observed that homologous sequences can vary widely in their nucleotide or amino-acid compositions, revealing that sequence evolution has changed importantly among lineages, and may therefore be most appropriately approached through non-homogeneous models. Several programs implementing such models have been developed, but they are limited in their possibilities: only a few particular models are available for likelihood optimization, and data sets cannot be easily generated using the resulting estimated parameters.

Results: We hereby present a general implementation of non-homogeneous models of substitutions. It is available as dedicated classes in the Bio++ libraries and can hence be used in any C++ program. Two programs that use these classes are also presented. The first one, Bio++ Maximum Likelihood (BppML), estimates parameters of any non-homogeneous model and the second one, Bio++ Sequence Generator (BppSeqGen), simulates the evolution of sequences from these models. These programs allow the user to describe non-homogeneous models through a property file with a simple yet powerful syntax, without any programming required.

Conclusion: We show that the general implementation introduced here can accommodate virtually any type of non-homogeneous models of sequence evolution, including heterotachous ones, while being computer efficient. We furthermore illustrate the use of such general models for parametric bootstrapping, using tests of non-homogeneity applied to an already published ribosomal RNA data set.

Background

In phylogenetics, simulations have been widely used to study the robustness of inference methods [1] and have been involved in parametric bootstrapping [2]. For instance, simulations have shown that maximum likelihood methods often more accurately reconstructed the

evolution of an alignment than distance or parsimony methods [3,4], but could also fail in conditions where compositional biases (a condition here referred to as non-homogeneity) or rate heterogeneity along branches (a phenomenon named heterotachy, [5]) were too intense [6-8]. Similarly, simulations have been used to compare topologies with respect to an alignment [9], or to assess the fit of a model to a particular data set [10-13]. In this last case, a model has a good fit to a particular data set if the alignments it generates have properties similar to the properties of the real alignment. Both for investigating reconstruction methods and for parametric bootstrapping, it is highly desirable that simulation methods model as precisely as possible the conditions that shaped biological sequences through evolution. However, widely-used simulation programs cannot be easily tuned to precisely reproduce the peculiar evolution of a particular data set. Noticeably, non-homogeneity cannot be simulated by Seq-Gen [14] or PAML [15], even if these phenomena are all known to affect the evolution of many data sets [5,16-20].

The ability to estimate parameters of sequence evolution with realistic models, and then computationally evolve sequences using these fitted parameters is crucial to better characterize the behavior of reconstruction methods in realistic settings.

Here we introduce extensions to the Bio++ package [21] that permit first to estimate parameters of evolution on a specific data set in a maximum likelihood framework, and second to simulate the evolution of sequences using these estimated parameters. Importantly, nearly any combination of non-homogeneous (including non-stationary models) and heterotachous models of evolution can be fitted to data, so that simulations may mimic very precisely the evolution of a data set. Such a flexibility should enable one to probe how robust methods of phylogenetic tree or ancestral state reconstruction are to more realistic evolutionary conditions. Moreover, it offers the possibility to compare a large variety of models by assessing through parametric bootstrapping their respective ability to reproduce a given characteristic of interest, measured on a real data set.

Implementation

Molecular phylogenetic methods are used by a wide range of biologists, from bioinformaticians willing to characterize and improve models of sequence evolution to molecular biologists trying to grasp the particular evolutionary history of their gene of interest. These different types of users have different needs: the former may benefit from easy-to-assemble, high-level object-oriented code to conduct phylogenetic analysis, while the latter likes user-friendly interfaces. However, both demand programs able to run the most recent models of evolution. The newly introduced extensions are available in two flavors that might fit different users' needs: (i) as classes in the Bio++ phylogenetic library, including a special class called SubstitutionModelSet which implements the relationships between models, parameters and branches, and (ii) through the BppML and BppSeqGen programs, which can respectively adjust these models to a data set and simulate data from these models. These programs share a common syntax for model specification and are hence fully interoperational and easy-to-use.

The SubstitutionModelSet class

The Bio++ libraries [21] provide data structures and algorithms dedicated to analysis of nucleotide, codon and amino acid sequences, phylogenetics and molecular evolution, and are designed in an object-oriented way. These include classes for storing phylogenetic trees, computing likelihood under various models of substitution, and estimating parameters. The likelihood classes take as input a phylogenetic tree and a substitution model, and were extended to allow the computation under non-homogeneous models (figure 1). This support is achieved through the addition of parameters for the rooting of the tree, since the likelihood may not be independent of the root position with a non-homogeneous model [6], and through a new class named SubstitutionModelSet. The SubstitutionModelSet class essentially associates a substitution model with each branch of the phylogenetic tree, and links each substitution model to a list of corresponding parameters (figure 2). It also provides a series of methods for the developer to set up the general model, to assign parameters to substitution models and substitution models to branches.

Substitution models can be totally independent of each other, or can share any number of parameters. Virtually any non-homogeneous model can thus be set up, provided the alignment is not a mix of nucleotide, amino-acid or codon sequences. All models available in Bio++ can be used with this class (*e.g.* K80, T92, HKY85, GTR, JTT92, etc), including heterotachous models (Galtier's model [22] and Tuffley and Steel's model [23]) and any rates across sites model (*i.e.* Gamma and Gamma + invariant distributions). The developer can also use the SubstitutionModelSet class with his own substitution model through the Bio++ SubstitutionModel interface. The SubstitutionModelSet class can be used in conjunction with other Bio++ classes to reconstruct ancestral states or to map substitutions, and hence allows to perform these analyses in the general non-homogeneous case.

Estimating parameters

Estimation of numerical parameters is performed using a modified Newton-Raphson optimization algorithm,

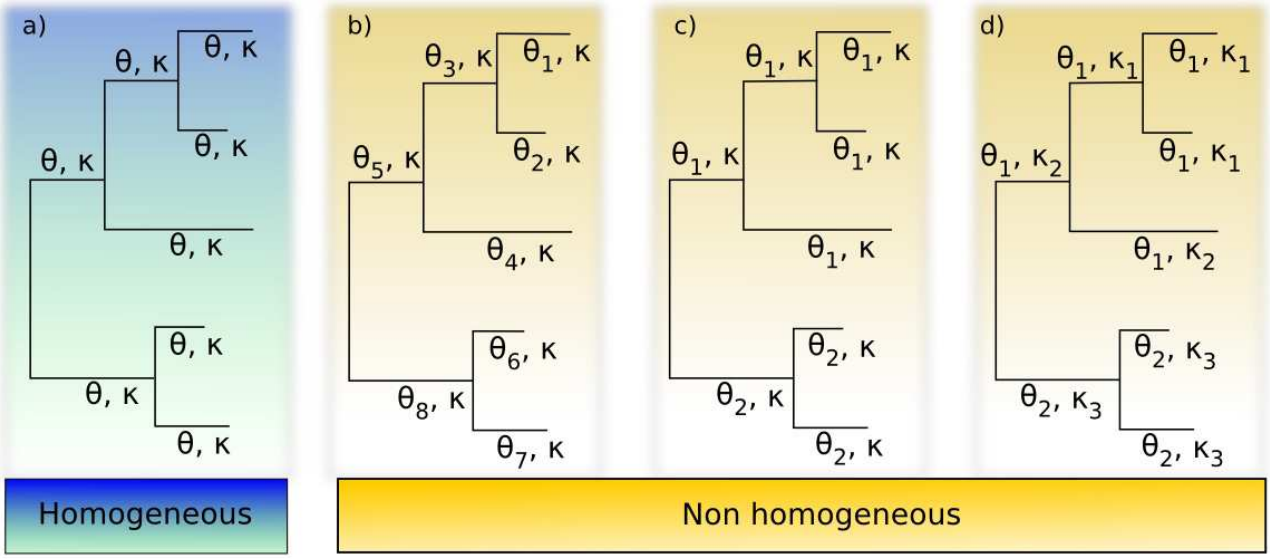


Figure 1
General non-homogeneous model of substitution. The substitution model depicted here is Tamura's 1992 model of substitution, which contains two parameters: κ, the transitions/transversions ratio and θ the equilibrium G+C content. In the homogeneous case, θ and κ are constant over the tree (case 'a'). In Galtier and Gouy's 1998 model, κ is constant over the tree and one distinct θ is allowed per branch (case 'b'). Between these two extrema lay models with certain branches, but not all, sharing a common value of θ (case 'c'). In the most general case 'd', there are two sets of parameters, one for κ and another for θ, that are shared by the branches of the tree.

commonly used in phylogenetics [4,24,25], and therefore requires computing derivatives with respect to parameters of the model. Because the use of the cross derivatives leads to numerical instabilities in the optimization (Nicolas Galtier, personal communication), they are set to zero in the Hessian matrix. Derivatives regarding branch lengths are computed analytically, whereas derivatives regarding the rates across sites distribution are computed numerically. Although the substitution model derivatives can be computed analytically in the homogeneous case as well as in Galtier and Gouy's model, they are difficult to compute analytically in the more general case, and are consequently computed numerically in Bio++. To prevent convergence issues due to erroneous derivative values we use, in the last optimization steps, Powell's multi-dimensions algorithm, which does not rely on parameter derivatives [26].

A general file format to describe non-homogeneous models

We introduced a new user-intuitive property file format to describe non-homogeneous substitution models. This format is an extension of PAML or NHML property file formats, and uses a syntax of the kind

property_name = property_value

A parser that automatically instantiates the appropriate SubstitutionModelSet object is included in the Bio++ libraries and is used by all programs in the Bio++ programs suite. Moreover, the same format is used for the input file of the programs and for their output, so that the output of one program (*e.g.* which adjusts a model to real data) can easily be used as the input of another one (*e.g.* which simulates data from a model). Figure 3 shows how the models in figure 1 are coded using this format. The core part of the description is the "model" property, which is associated to one or several nodes of the phylogenetic tree through node identifiers. These node identifiers can be obtained from the programs in the Bio++ program suite, or set by the user in his own program.

The BppML and BppSeqGen programs

Parameter estimation and simulation procedures are available as dedicated classes in the Bio++ phylogenetic library, and can hence be used in any C++ program. However, for users who would rely on appropriate software rather than program their own tools, the Bio++ program suite was designed. These programs, including BppML (for Bio++ Maximum Likelihood) and BppSeqGen (Bio++ Sequence Generator) are command line driven and fully parametrized using property files, as introduced above.

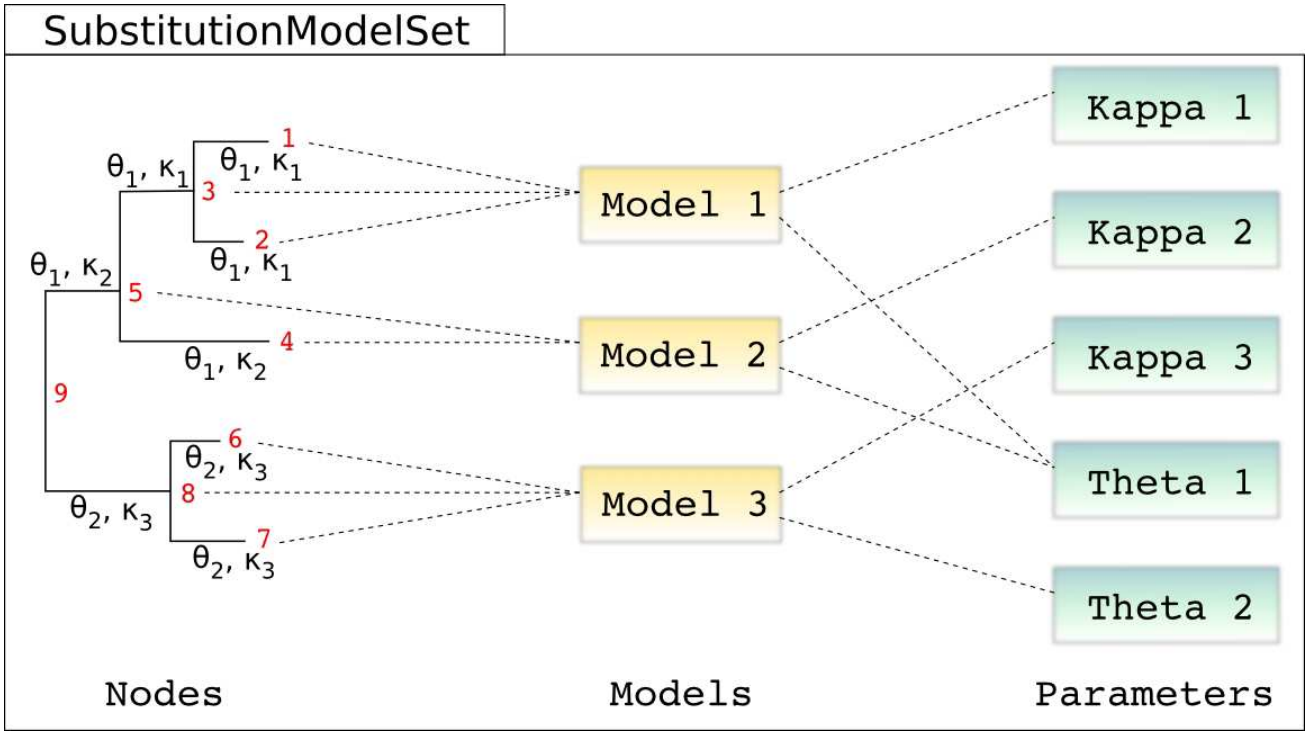


Figure 2
Relations between branches, models and parameters. In the general non-homogeneous case, model parameters are shared by different branches across the tree. These parameters are part of branch-specific substitution models, which specify branch-wise probabilities of replacement between states. Branches are here defined according to their rightmost node. The SubstitutionModelSet class stores dependencies between nodes, models and parameters.

They can thus easily be pipelined with scripting languages as bash, python or perl. In addition to the BppML and BppSeqGen programs, the Bio++ program suite also contains programs for distance-based phylogenetic reconstruction, sequence file format conversion and tree manipulation.

Results and Discussion

Our new general non-homogeneous model implementation was applied to Boussau and Gouy's data set of concatenated small and large subunit ribosomal RNA sequences and tree [6]. This data set contains 92 sequences and 527 complete sites. We first compare computation time, memory usage and parameter estimation for various models and software. We then show how the general non-homogeneous model introduced here can be used to study model fit through parametric bootstrapping.

In this section, we use the following model notations:

H Homogeneous model, using a Tamura 1992 substitution model [27].

NH1 One-theta-per-branch non-homogeneous model [24]. This model uses Tamura's 1992 substitution model, with one θ (equilibrium G+C content) per branch in the tree, whereas κ (transitions/transversions ratio) is shared by all branches.

NH2 One-theta-per-kingdom non-homogeneous model. In this general model, we allowed each kingdom (Bacteria, Eukaryotes or Archaea) to have its own equilibrium G+C content, while sharing the same transitions/transversions ratio.

NH3 Same as NH2, but in addition the (hyper)thermophilic Bacteria on one hand, and the eukaryote G+C-rich genus Giardia on the other hand were allowed to have their own equilibrium G+C content.

NH4 One-kappa-per-branch non-homogeneous model. This model has one κ per branch in the tree, whereas θ is shared by all branches.

Performance

We compared the likelihood of our implementation with the NHML [22,24] and [nh]PhyML [4,6] programs (see

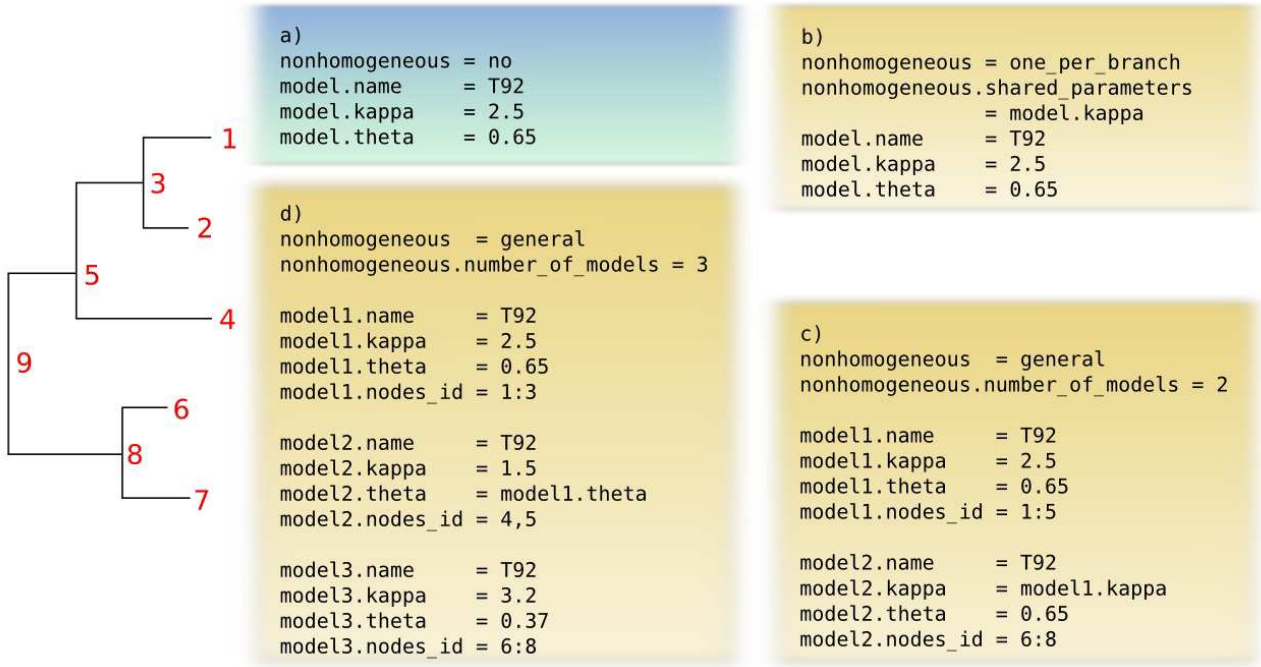


Figure 3
Model specification in BppML and BppSeqGen. A general file format is introduced to allow for the user-friendly description of virtually any non-homogeneous model. The tree shows the nodes identifiers, which can be obtained from the programs or defined by the user in its own program. Each case presented here corresponds to a particular model in figure 1, and was labeled accordingly. Each parameter can be fixed to a specific value or optimized with BppML.

table 1 and Additional file 1). Several models have been tested: Kimura two parameters (K80) for the homogeneous case, and Tamura 1992 (T92) derived models for the non-homogeneous cases, with constant rate, Gamma distributed rates (4 classes), Gamma (4 classes) + invariant and Galtier's 2001 site-specific rate variation model (covarion-like). On all tested models, the optimization algorithm in Bio++, while using numerical derivatives, leads to similar or better likelihood values than other programs, although at the price of an increase in computational time. However this increase is not sufficient to prevent the use of complex models on data sets of usual sizes, as it takes a little bit more than an hour and a quarter to optimize parameters with the richest models on a data set containing 92 sequences. It is also noteworthy that the Bio++ implementation requires less memory than other programs. This is partly explained by differences in the algorithms used to compute the likelihood [28]. The PhyML programs, including nhPhyML, use a double-recursive algorithm [6], which saves a lot of computation when exploring the space of tree topologies but results in a three fold increase in memory usage compared to the simple-recursive algorithm. Because no tree space exploration was involved, BppML computations used the simple-recursive algorithm. If desired, however, Bio++ also offers the double-recursive algorithm.

The convergence of the optimization algorithm was assessed by two methods, using the NH3 model. First, we used 100 distinct randomly chosen initial sets of parameter values and the RNA data set (see methods). We found that the estimated values obtained in each run were the same for all parameters up to the 5th decimal. Second, we simulated 100 data sets using the NH3 model with a Gamma + invariant rate distribution, with parameter values estimated from the real data set and the same number of sites. These parameters were then re-estimated for each simulated data set using random initial values. The results are displayed on figure 4, and show that the parameter values are recovered without bias and with a good precision. The only exception is the proportion of invariant sites which is slightly overestimated. These results also validate the simulation procedure.

Example of application: parametric bootstrap and Bowker's test for non-homogeneity

As most phylogenetic reconstruction models are homogeneous, they do not properly model the evolution of

Table 1: Comparison of the NHML, (NH)PhyML and BppML programs. Likelihood: - log (likelihood) of the optimized parameters, with a fixed tree topology.

Likelihood												
Rate	Constant			Γ(4)			Γ(4) + *I*			Covarion		
Model	H	NH1	NH3	H	NH1	NH3	H	NH1	NH3	H	NH1	NH3
NHML	15307	15034	--	14145	13828	--	--	--	--	13750	**13397**	--
PhyML	**15187**	15011	--	**14141**	13824	--	**14128**	--	--	--	--	--
BppML	15187	**14920**	15109	14141	**13821**	14029	14128	13810	14018	**13747**	13399	13615
Time												
Rate	Constant			Γ(4)			Γ(4) + *I*			Covarion		
Model	H	NH1	NH3	H	NH1	NH3	H	NH1	NH3	H	NH1	NH3
NHML	0:01:40	0:02:28	--	0:03:07	**00:02:13**	--	--	--	--	0:19:24	**0:19:09**	--
PhyML	**0:00:07**	**0:01:43**	--	**0:00:34**	00:02:29	--	**0:00:35**	--	--	--	--	--
BppML	0:00:27	0:11:57	0:01:12	0:00:47	00:35:46	0:00:48	0:01:01	0:29:40	0:01:38	**0:02:52**	1:14:32	0:14:27
Memory												
Rate	Constant			Γ(4)			Γ(4) + *I*			Covarion		
Model	H	NH1	NH3	H	NH1	NH3	H	NH1	NH3	H	NH1	NH3
NHML	16.38	20.48	--	55.30	65.54	--	--	--	--	55.30	65.54	--
PhyML	10.24	28.67	--	30.73	77.82	--	30.72	--	--	--	--	--
BppML	**08.19**	**08.19**	08.19	**14.34**	**14.34**	14.34	**14.34**	16.38	16.38	**12.29**	**14.34**	12.29

Time is shown as hours:minutes:seconds. Numbers in bold font correspond to the best performance for each comparison. Memory corresponds to the maximum memory usage during the program execution in megabytes. H: homogeneous case, with a K80 substitution model, NH1: theta per branch model, with a T92 substitution model, NH3: clade-specific and G+C-rich species theta model, see methods. The PhyML program was used for the H model, and nhPhyML for the NH1 model.

homologous sequences that vary widely in their compositions. Analyzing compositionally heterogeneous data sets with homogeneous models of sequence evolution may therefore lead to incorrect inferences, provided the heterogeneity is large enough. Several tests have been developed to assess the amount of heterogeneity present in a data set (see [29] for a review).

Estimating the amount of compositional heterogeneity in a data set

Most commonly, a matrix is assembled that contains compositions in all characters for all sequences, and this matrix is analyzed through χ^2 statistics [29]. However, this approach usually does not distinguish between constant and variable sites, and therefore may underestimate the true amount of heterogeneity in a data set [29].

Recently, Ababneh et al. [30] re-introduced Bowker's pairwise test [31] for symmetry. Given two aligned sequences S_1 and S_2 on a given alphabet of size n and characters $x_{\{1,2...n\}}$, it compares the numbers of substitutions between x_i in S_1 and x_j in S_2, $\{i,j\} \in [1:n]$, with the numbers of substitutions between x_j in S_1 and x_i in S_2. If these pairs of numbers are equal for all $\{i, j\} \in [1:n]$, the two sequences may have evolved according to two identical processes. Otherwise, the two processes were necessarily different.

Bowker's test therefore permits to assess whether compositional differences have accumulated between two sequences through non-homogeneous evolution. To apply it to more than two sequences, Rodriguez-Ezpeleta et al. [32] computed all pairwise Bowker's tests in their alignment and computed the median value; one could also have counted the number of Bowker's tests that are significant at a 5% threshold according to a χ^2 table.

However, none of these tests permit to estimate if the amount of heterogeneity that they detect in a given data

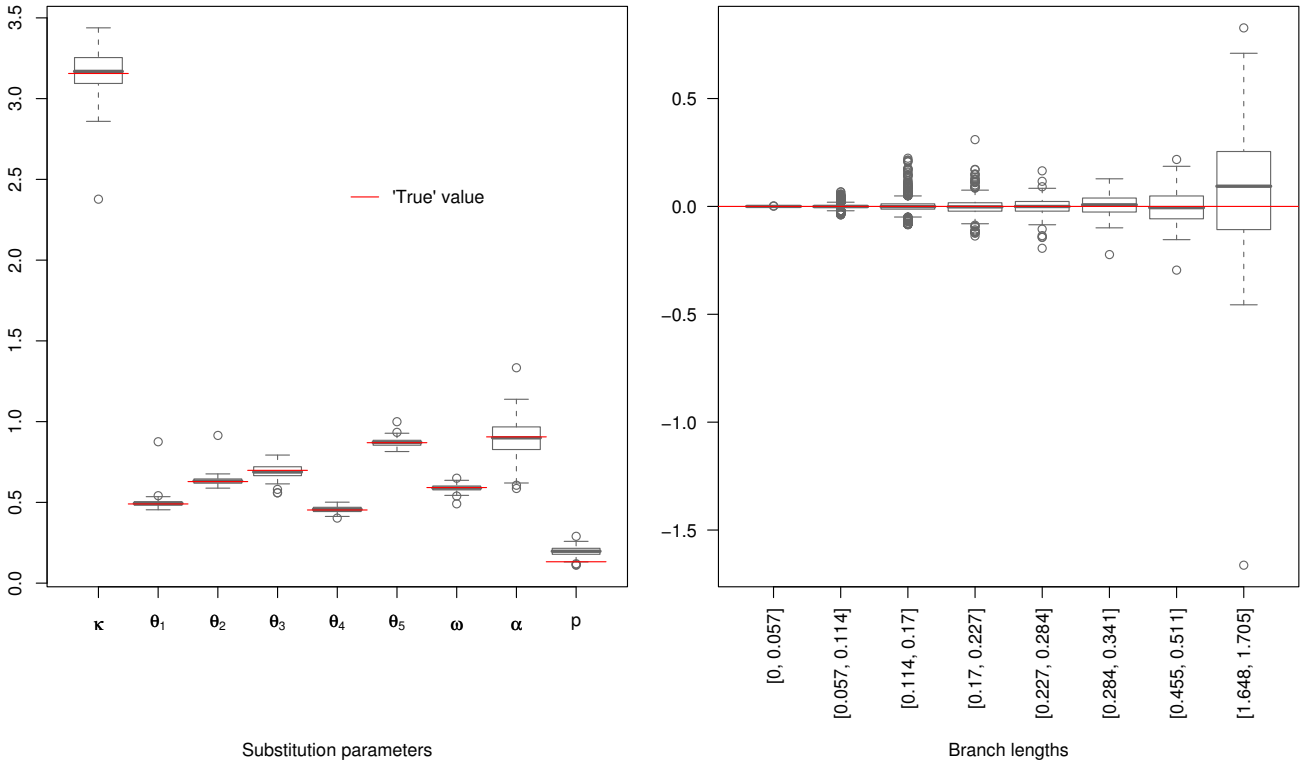


Figure 4
Assessing parameter estimation using simulations. Left: boxes show the median and quartiles of the distribution of parameter estimates for 100 simulations. The 'true' value used in the simulation is shown in red. Right: boxes show the distribution of the bias (estimated value – real value), as a function of the (pooled) real values of the branch length. θ_i: GC content, ω GC content at root, α : shape of the Gamma distribution of rates across sites, p proportion of invariant sites (see text for details on the model used).

set is sufficient to bias inferences made using homogeneous models, although this is likely the question an average user would like to answer.

Assessment of the fit of evolutionary models with respect to compositional heterogeneity

Here, we describe a method to reveal the ability of evolutionary models to account for the compositional heterogeneity in a sequence alignment, which we measure using the median of all Bowker's pairwise statistics, or the number of significant Bowker's pairwise tests (in the following, we note the measure of compositional heterogeneity h). This method is tree-based, and uses parametric bootstrapping [10-12]. In this respect, it is similar to the method recently introduced in [13] in the Bayesian setting. Our approach requires 5 steps to estimate the fit of a model M to a data set D.

1. Compute the compositional heterogeneity measure h for the data set D.

2. Estimate the parameters of model M based on the data set D according to the Maximum Likelihood criterion.

3. Simulate a large number of data sets D' using the model M previously estimated.

4. Compute the compositional heterogeneity measure h' for each alignment D'.

5. Compare the measure h obtained on data set D to measures h' obtained on data sets D'. If h is outside 95% of the distribution of h', the model does not properly reproduce the heterogeneity of data set D.

Using such an approach, any model can be compared with others with respect to their ability to handle the compositional heterogeneity of a given data set: the closest the distribution of h' is from h, the highest is the fit. Ideally, the distribution of measures h' obtained on the parametric bootstrap replicates of a good model should be cen-

tered around the value obtained for the real alignment *h*, with a very low variance. If one neglects potential problems linked with over-parametrization, the inferences of the best model should be preferentially trusted compared to a model that fails to account for an important feature of a data set. Overall, our approach can be used for model selection, although contrary to criteria such as AIC or BIC [28] this approach does not take into account the number of parameters; more importantly, it can also be used for estimating model adequacy.

Application to an rRNA data set

Our approach to assess the composition-wise fit of evolutionary models to a data set was applied to an alignment containing ribosomal RNA sequences from Archaea, Bacteria and Eukaryotes [6]. First, several homogeneous and non-homogeneous models were fitted to the data set, using a Tamura 1992 model of substitution with a four classes Gamma + invariant distribution of rates across sites. Then, 10,000 artificial data sets were simulated in each case using these estimated parameters. Eventually, the real data set and the simulated data sets were compared with respect to their compositional heterogeneity: models able to simulate data sets with similar amounts of heterogeneity as the real data set appropriately account for this specific aspect of the data.

Results are shown in figure 5 and table 2. Both the number of significant Bowker's tests and the median of their values give similar results. For instance, both indices find that the real data set shows significantly more heterogeneity than the distributions of data sets simulated under the homogeneous model of sequence evolution (p-value = 0.0008 for the number of significant pairwise tests and p-value = 0.0028 for the median). The homogeneous model therefore lacks parameters useful to account for this particular feature of the data. Allowing different transition/transversion rates for each branch as in model NH4 does not solve this problem, as the obtained bootstrapped distribution also significantly underestimates the heterogeneity in the real data (p-value = 0.0015 and p-value = 0.0047, respectively). It is noteworthy, however, that the likelihood ratio test finds that this model describes the data significantly better than the homogeneous one, whereas the AIC and BIC criteria do not. On the contrary, the NH1 model simulated sequences distribution surrounds the value obtained on the real data set (p-value > 0.7 in both cases). This suggests that Galtier and Gouy's modeling [24] properly accounts for the heterogeneity in rRNA data sets, and that there may be no point in using more parameter-rich models such as Yang and Roberts' [33] on these molecules. The results even suggest that NH1 might be slightly prone to over-estimating the amount of heterogeneity. For instance, the median Bowker's test value for simulated data sets are most often higher than the value obtained on the real data set. NH1's behavior may be explained by over-parametrization: it is likely that during sequence evolution, not all branches witnessed significant shifts in mutational parameters or selection pressures. To investigate further the impact of the number of parameters on model fit, two other models were tested: NH2, in which different equilibrium G+C contents are associated to each kingdom, and NH3, which further adds two equilibrium G+C contents, one for the hyperthermophilic (G+C rich) Bacteria, and one for the G+C rich Eukaryote Giardia. Hyperthermophilic (G+C rich) Archaea were not considered separately from the others as nearly all Archaea in our data set were thermophilic or hyperthermophilic. The NH2 model seems to lack useful parameters to properly account for the heterogeneity in the real data set, as its simulated data sets are less heterogeneous than the real one (p-value = 0.0040 for the number of pairwise tests, and 0.0141 for the median). The NH3 model improves upon NH2 as its bootstrapped distribution is more centered upon the observed value, which is no longer rejected (p-value = 0.14 and 0.27). However, the observed value is still on the right side of the null-distribution, and it is very likely that the correct parametrization lays between NH1, too rich with its 182 equilibrium G+C contents, and NH3, maybe too poor with its 5 equilibrium G+C contents. However, as NH3 provides a fit nearly as good as NH1 with a much lower amount of parameters, the best model may well have less than a dozen equilibrium G+C contents. Interestingly, Bowker's tests are in agreement with the Bayesian information criterion (BIC, see table 2) and favor the NH3 model. Conversely, Akaike's information criterion (AIC) and the likelihood ratio test (LRT) favor the more parameter-rich model NH1. Obviously, although a few works already addressed this issue in the Bayesian framework [11,13,34], automatic ways to explore and choose among heterogeneous models in a maximum likelihood framework are much needed. All the tools required for such a project are now available in the Bio++ libraries.

Conclusion

Bio++ is a growing set of libraries designed for sequence, phylogenetic and molecular evolution analyzes. In this article extensions allowing to implement a wide variety of non-homogeneous models of sequence evolution were introduced. Combined with support for rates across sites and heterotachous models of evolution, and with routines for optimizing parameters and tree topology in the maximum likelihood framework, they provide a comprehensive platform for phylogenetic studies, either for bioinformaticians willing to develop their own software, or for biologists characterizing the evolution of a particular set of sequences using the BppML and BppSegGen programs. Whilst being a generalist program implementing a large variety of models, BppML was shown to be of a sim-

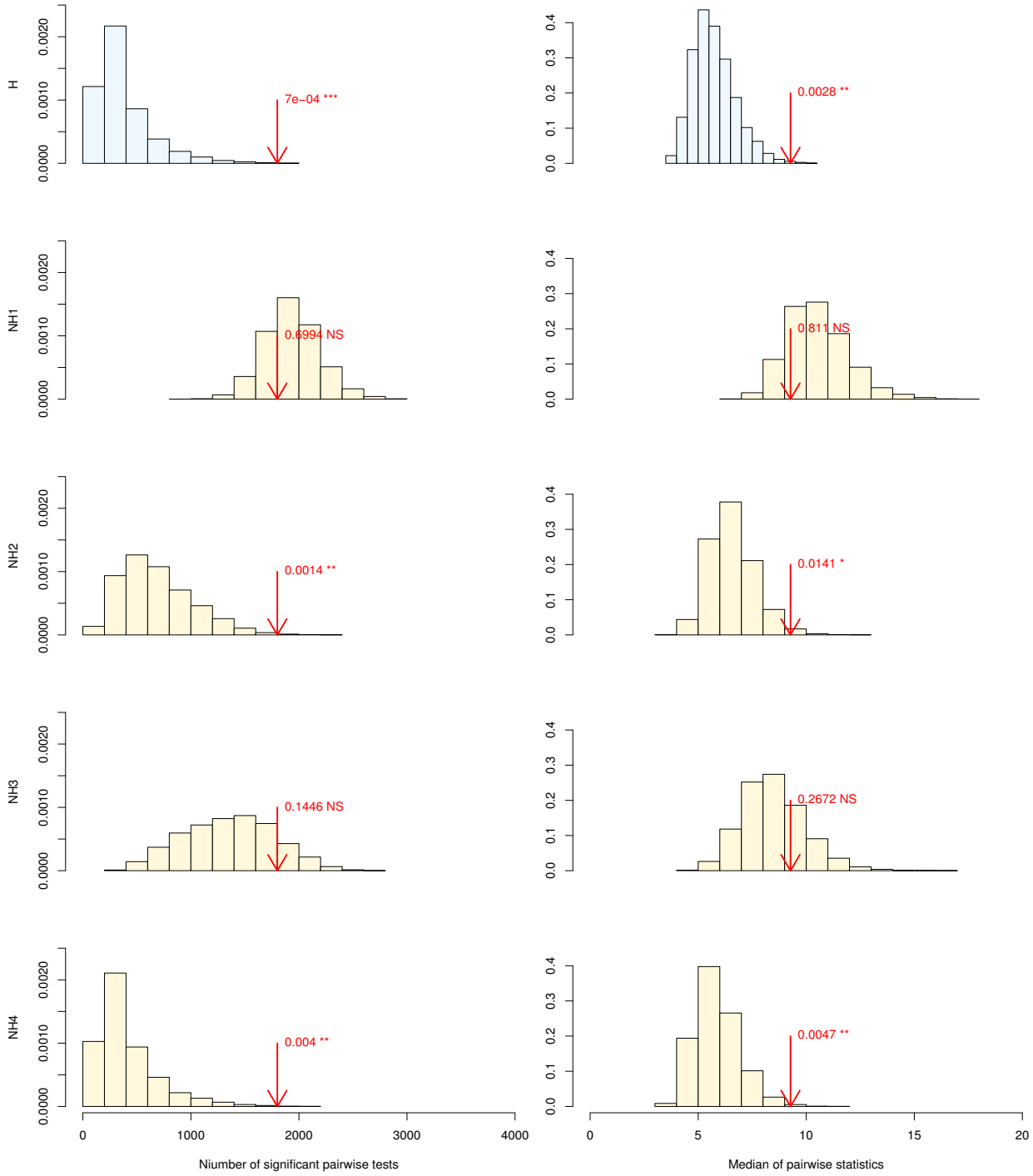


Figure 5

Distributions of the Bowker's test statistics under various models. First column: number of pairwise tests significant at the 5% level. Second column: median of the pairwise statistics. First row: homogeneous model (H). Second row: one theta per branch non-homogeneous model (NH1). Third row: 3 thetas non-homogeneous model (NH2). Fourth row: 5 thetas non-homogeneous model (NH3). Fifth row: one kappa per branch non-homogeneous model (NH4). All models use the Tamura 1992 substitution model with a 4-classes discrete Gamma + invariant rate distribution. The arrows indicate the observed values from the real data set and the resulting p-values.

Table 2: Model comparisons.

Model	lnL	*k*	LRT			AIC	BIC	Bowker	
			H	NH2	NH3			# tests	median
H	-14110.628293	185				28591.26	29380.69	0.0008	0.0028
NH1	-13810.371502	368	600.51	556.74	416.97	**28356.74**	29927.07	0.7010	0.8110
NH2	-14088.739682	189	43.78			28555.48	29361.98	0.0040	0.0141
NH3	-14018.854234	191	183.55	139.77		28419.71	**29234.74**	0.1448	0.2672
NH4	-13970.841467	368	279.57			28677.68	30248.01	0.0015	0.0047

Comparison of the various non-homogeneous models with the homogeneous case, using different criteria. *k* is the number of parameters and *lnL* is the log likelihood of each model. The Akaike's information criterion (AIC) of each model is defined as $2k - 2 \cdot lnL$, and the lowest value, corresponding to the best model according to this criterion is in bold font. The Bayesian information criterion (BIC) is computed as $k \cdot \ln(n) - 2 \cdot lnL$, $n = 527$ being the number of observations. The lowest value is in bold font. The likelihood ratio test (LRT) allows to compare nested models only, and is defined as minus two times the logarithm of the ratio of likelihoods. All LRT are significant at the 0.1% level. This ratio follows a χ^2 distribution with the number of additional parameters as the degrees of freedom. The last two columns show the p-values of the two Bowker's test introduced in this paper.

ilar quality as programs dedicated to particular homogeneous or non-homogeneous models of evolution, achieving higher likelihood scores with smaller memory requirements while conserving reasonable running-times. Its joint use with BppSeqGen permits to precisely study the evolution of a particular data set through parametric bootstrapping, and may be used to generate realistic artificial data sets to study the robustness of phylogenetic reconstruction methods in the presence of heterogeneity and heterotachy. Further developments may involve methods to optimize the number of models necessary to account for the heterogeneity in a data set, or methods to explore the space of tree topologies with a broad range of non-homogeneous models of sequence evolution.

Methods

Data and phylogeny reconstruction

RNA sequences from the small and the large subunit of the ribosome were aligned and concatenated. Sequences coming from 22 Archaea, 34 Bacteria and 36 Eukaryotes were selected to yield a data set containing 92 sequences and 527 complete sites, with G+C contents ranging from 43% to 71%. A phylogenetic tree was built with nhPhyML [6]. For additional information, please refer to [6].

Comparing likelihood optimizations

The NHML, (NH)PhyML and BppML programs were used to compare optimization performances. The programs were run on the data set from [6], after all columns in the alignment containing at least either a gap or an unknown character had been removed. The phylogenetic tree from [6] was used as a fixed topology, and the branch lengths used as initial values for the optimization. To allow the comparison between the three programs, the Kimura two parameters model of substitution [35] was used for homogeneous models and models derived from Tamura's 1992 model [27] for non-homogeneous models. Initial values were set to 1 and 0.5 for the κ and θ parameters respectively. A Gamma (4 classes) + invariant rates across sites distribution was also tested, with initial value set to 0.5 for the Gamma shape parameter, and 0.2 for the proportion of invariants. Galtier's 2001 [22] heterotachous model was also tested, with 4 rate classes, initial values of the shape parameter set to 0.5, and initial value of the rate change parameter set to 0.5. The precision in the optimization algorithm was set to 0.000001 for the three programs. The total length of execution was corrected according to the average CPU usage, and the memory usage corresponds to the maximum reached during program execution, as reported by the Unix "top" command. All calculations were performed on a 64 bits Intel(R) Core(TM)2 Duo, CPU 2.66 GHz.

Assessing the convergence of the optimization procedure

Different initial values were used as initial guesses for the optimization algorithm. The GC frequencies and the proportion of invariant sites were chosen randomly from a uniform distribution between 0 and 1. The transitions/transversions ratio and the alpha parameter of the rate distribution were picked from a [0, 5] and [0.2, 2] uniform distributions, respectively. Branch lengths were taken from a uniform distribution between 0 and 0.1.

Computing p-values for Bowker tests

Alignment-wise tests for non-homogeneity were performed using two types of statistics:

- The number of 5% significant pairwise tests,
- The median of pairwise statistics.

In both cases, the global p-value was computed as

$$\text{p} - \text{value} = \frac{N_2 + 1}{N_1 + 1}, \quad (1)$$

where N_1 is the number of simulations performed under the null model, and N_2 is the number of values of the sta-

tistic in the simulations that were greater or equal to the observed one, measured from the real data set. In this study, N_1 was set to 10,000.

Program source code for performing Bowker's test is provided as Additional file 2. The data and scripts to run the analyses are in Additional file 3.

Availability and requirements

Project name: The Bio++ libraries (version 1.6) and programs suite (version 1.0).

Project home page: http://kimura.univ-montp2.fr/BioPP and http://home.gna.org/bppsuite

Operating systems: Any platform with a C++ compiler and supporting the Standard Template Library

Programming language: C++

Other requirements: The C++ Standard Template Library

License: The CeCILL free software license (GNU compatible)

Authors' contributions

BB and JD designed the method, implemented the software and wrote the article. JD ran the analyses.

Additional material

Additional file 1

Detailed results of model comparison. *OpenDocument spreadsheet (.ods) file containing detailed results from table 1, with parameter estimates obtained.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-255-S1.ods]

Additional file 2

Program to compute Bowker's test. *Zip archive containing the C++ program used to compute Bowker's test.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-255-S2.zip]

Additional file 3

Data set, tree and scripts for running Bowker's tests. *Zip archive containing the sequence alignment and phylogenetic tree used, together with scripts for running the tests presented in this article.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-8-255-S3.zip]

Acknowledgements

The authors would like to thank Manolo Gouy, Nicolas Galtier, Mathieu Emily and Matthew Spencer for helpful comments on this manuscript.

References

1. Williams PD, Pollock DD, Blackburne BP, Goldstein RA: **Assessing the accuracy of ancestral protein reconstruction methods.** *PLoS Comput Biol* 2006, **2:**e69-e69.
2. Goldman N: **Statistical tests of models of DNA substitution.** *J Mol Evol* 1993, **36:**182-198.
3. Kuhner MK, Felsenstein J: **A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.** *Mol Biol Evol* 1994, **11:**459-468.
4. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696-704.
5. Lopez P, Casane D, Philippe H: **Heterotachy, an important process of protein evolution.** *Mol Biol Evol* 2002, **19:**1-7.
6. Boussau B, Gouy M: **Efficient likelihood computations with nonreversible models of evolution.** *Syst Biol* 2006, **55:**756-768.
7. Kolaczkowski B, Thornton JW: **Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous.** *Nature* 2004, **431:**980-984.
8. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F: **Heterotachy and long-branch attraction in phylogenetics.** *BMC Evol Biol* 2005, **5:**50-50.
9. Goldman N, Anderson JP, Rodrigo AG: **Likelihood-based tests of topologies in phylogenetics.** *Syst Biol* 2000, **49:**652-670.
10. Bollback JP: **Bayesian model adequacy and choice in phylogenetics.** *Mol Biol Evol* 2002, **19:**1171-1180.
11. Foster PG: **Modeling compositional heterogeneity.** *Syst Biol* 2004, **53:**485-495.
12. Lartillot N, Brinkmann H, Philippe H: **Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model.** *BMC Evol Biol* 2007, **7(Suppl 1):**S4-S4.
13. Blanquart S, Lartillot N: **A Site- and Time-Heterogeneous Model of Amino-Acid Replacement.** *Mol Biol Evol* 2008.
14. Rambaut A, Grassly NC: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Cabios* 1997, **13:**235-238.
15. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586-1591.
16. Sueoka N: **On the genetic basis of variation and heterogeneity of DNA base composition.** *Proc Natl Acad Sci USA* 1962, **48:**582-592.
17. Galtier N, Lobry JR: **Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes.** *J Mol Evol* 1997, **44:**632-636.
18. Foster PG, Jermiin LS, Hickey DA: **Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria.** *J Mol Evol* 1997, **44:**282-288.
19. Zeldovich KB, Berezovsky IN, Shakhnovich EI: **Protein and DNA sequence determinants of thermophilic adaptation.** *PLoS Comput Biol* 2007, **3:**e5-e5.
20. Wang HC, Spencer M, Susko E, Roger AJ: **Testing for covarion-like evolution in protein sequences.** *Mol Biol Evol* 2007, **24:**294-305.
21. Dutheil J, Gaillard S, Bazin E, Glémin S, Ranwez V, Galtier N, Belkhir K: **Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics.** *BMC Bioinformatics* 2006, **7:**188-188.
22. Galtier N: **Maximum-likelihood phylogenetic analysis under a covarion-like model.** *Mol Biol Evol* 2001, **18:**866-873.
23. Tuffley C, Steel M: **Modeling the covarion hypothesis of nucleotide substitution.** *Math Biosci* 1998, **147:**63-91.
24. Galtier N, Gouy M: **Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis.** *Mol Biol Evol* 1998, **15:**871-879.
25. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6.** *Distributed by the author* 2005.
26. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C. The Art of Scientific Computing* second edition. Cambridge University Press; 1992.

27. Tamura K: **The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA.** *Mol Biol Evol* 1992, **9:**814-825.
28. Felsenstein J: *Inferring Phylogenies* Sinauer Associates, Inc; 2004.
29. Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW: **The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated.** *Syst Biol* 2004, **53:**638-643.
30. Ababneh F, Jermiin LS, Ma C, Robinson J: **Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences.** *Bioinformatics* 2006, **22:**1225-1231.
31. Bowker A: **A test for symmetry in contingency tables.** *J Am Stat Assoc* 1948, **43:**572-574.
32. Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H: **Detecting and overcoming systematic errors in genome-scale phylogenies.** *Syst Biol* 2007, **56:**389-399.
33. Yang Z, Roberts D: **On the Use of Nucleic Acid Sequences to Infer Branchings in the Tree of Life.** *Mol Biol Evol* 1995, **12:**451-458.
34. Blanquart S, Lartillot N: **A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution.** *Mol Biol Evol* 2006, **23:**2058-2071.
35. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16:**111-120.

8

Coping with Heterogeneous Evolutionary Roads in a Single Gene

Article 3 showed that composition heterogeneity was not the only problem that a phylogeneticist had to confront to build a phylogenetic tree. Lateral gene transfer notably was a very important difficulty, that gets even tougher when one accepts that not only whole genes are exchanged between species, but that so can be parts of genes, through a process named recombination. In such circumstances, one gene may have several different histories.
In this article, we developed models to estimate the histories that may lie hidden in gene sequences. If many genes have undergone recombination events, it might be important to use such models to reconstruct species phylogenies.

This article has not been submitted yet.

A Mixture Model and a Hidden Markov Model to Detect Recombination

Bastien Boussau*, Laurent Guéguen and Manolo Gouy

December 1, 2008

*Corresponding author
Université de Lyon ; université Lyon 1 ; CNRS ; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France.

E-mail: boussau@biomserv.univ-lyon1.fr

key words: molecular phylogeny; recombination; maximum likelihood; PhyML;

Abstract

Homologous recombination is a pervasive biological process that affects sequences in all living organisms and viruses. In the presence of recombination, the evolutionary history of an alignment of homologous sequences cannot be properly depicted by a single bifurcating tree: some sites have evolved along a specific phylogenetic tree, others have followed another path. Methods available to analyse recombination in sequences usually involve a painstaking analysis of the alignment through sliding-windows, or are particularly demanding in computational resources, and are often limited to nucleotidic sequences. In this article, we propose and implement a Mixture Model on trees and a phylogenetic Hidden Markov Model to reveal recombination breakpoints while searching for the various evolutionary histories that are present in the alignment. These models are sufficiently efficient to be applied to dozens of sequences, and can handle indifferently nucleotidic or proteic sequences. We estimate their accuracy on simulated sequences and test them on real data.

Introduction

Homologous recombination is a process through which genes descending from a same ancestor exchange parts of their sequence. Consequently, sequences having undergone recombination will display two different histories: one history for the non-recombining part of their sequence, and one history for the recombining part. If the recombining genes have been parts of different lineages long enough prior to this recombination event, the difference in the histories of the recombining and non-recombining parts of the gene may translate into topological incongruencies between their respective phylogenies.

If one applies classical phylogenetic methods onto an alignment that has undergone recombination, only one tree will be recovered, with no guarantee that this tree corresponds to the recombining part of the sequence, the non-recombining part, or any of these two. Several methods have been developped to try and detect recombination in alignments [1, 2]; such methods can therefore be used prior to phylogenetic analysis to see whether it is meaningful to describe the history of an alignment by a single bifurcating tree. In cases where no recombination has been detected, the subsequent analysis is classical phylogenetics. In cases where recombination has been detected, there are few methods available that can analyse an alignment and precisely predict both the recombination breakpoints and the evolutionary histories found in the alignment.

If we put aside methods based on sliding windows, that are painstaking and cannot precisely pinpoint the recombination breakpoints, two groups have proposed methods to unveil both the recombination positions and the phylogenetic trees. In 2000, Mcguire et al. [3], inspired by the work of Felsenstein and Churchill [4], proposed a method based on a hidden Markov model (HMM) in which the hidden states were the phylogenetic trees themselves. Therefore, a transition between the states ought to be a recombination breakpoint. However, this first attempt was prone to detecting recombination events where there was only rate heterogeneity. Husmeier subsequently built upon this model to deal with heterogeneities in site evolutionary rates [5] by superimposing another HMM whose states correspond to evolutionary rates: therefore two kinds of transitions are allowed along the alignment, a transition between topologies, indicative of recombination, and a transition between rates. Unfortunately, all these methods are computationaly demanding, and can only be applied in cases where the space of tree topologies is very limited, as all topologies need to be given *a priori*. Lastly, Kedzierska and Husmeier [6, 7] proposed a hybrid approach in which a sliding window is first applied to the alignment to build phylogenetic tree distributions along the alignment. Then, a HMM is run on the alignment, with its hidden states being the tree distributions themselves. This approach allows to handle a greater number of sequences than the previous ones, but is also probably less accurate in the detection of the breakpoints, because the topology distributions are built from small arbitrary windows, which may not correspond to the true recombination structure of the alignment.

In 2002, Suchard and co-workers [8] proposed a bayesian multiple-changepoint model to detect recombination, and further improved it by adding a second changepoint process to account for changes in the substitutional process [9, 10]. This sophisticated method however also suffers from its computational requirements. In fact, both this method and the ones of Husmeier, Wright and co-workers have been implemented to only deal with DNA sequences, and can not be used with large numbers of sequences.

However, the detection of recombination should not be limited to recently diverged sequences. When protein-coding sequences have diverged a long time ago, the nucleotide sequence may be saturated, so that it becomes mandatory to resort to amino-acid sequences. In such conditions, none of the previously described method can be used.

Most recently, Pond and co-workers developped GARD [11, 12], a software able to detect recombination with any type of alphabet. This program estimates the phylogenetic trees, the number of recombination breakpoints and their positions in a maximum likelihood framework. To do so, it tries different numbers of breakpoints, and for each number, uses a genetic algorithm to estimate the best breakpoint positions. During this procedure, phylogenetic trees are estimated with the Neighbor-joining algorithm [13], and the best number and positions of breakpoints are chosen according to the Akaike criterion. This considerable task can be achieved efficiently through

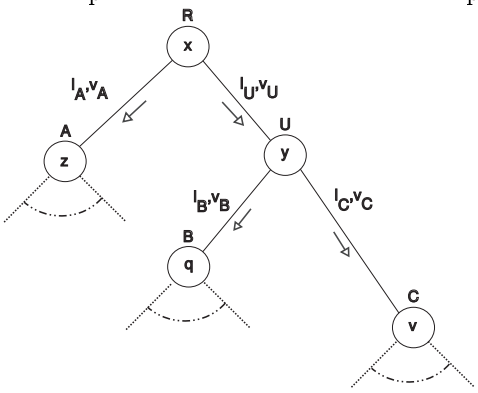
a parallelised architecture, which can be run on a cluster of computers.

In this article, we present two new methods to uncover the recombination structure of a protein or nucleotidic alignment, that can be easily and efficiently run on a desktop computer. The first method is based on a Mixture Model (MM), and the second is based on a phylogenetic Hidden Markov Model (Phylo-HMM). We begin by introducing the mathematics behind these models, shortly explain how these were implemented, and finally proceed to test them on both simulated and real alignments. We discuss the merits and limits of our methods and propose a few refinements.

Computing the Likelihood of a Single Tree

We first explain how one computes the likelihood of a phylogenetic tree [14] with nucleotidic or protein sequences using the following example (Fig. 1).

Figure 1: Example rooted tree for likelihood computation.



Most commonly, sites are supposed to evolve independently of each other: a site does not depend on its neighbors' states but only on its past state. As a consequence, the likelihood of a tree for a whole sequence is obtained by multiplying all the likelihoods obtained at single sites.

The likelihood $L_{s,\tau}$ of the tree τ given in Fig. 1 for a single site s is computed as follows:

$$
\begin{aligned}
L_{s,\tau} \quad = \quad & \sum_{g\in\Gamma} \frac{1}{|\Gamma|} \Bigg[\sum_{x\in\Omega} \Bigg[P(R=x) \\
\times \quad & \sum_{z\in\Omega} \left[P_{xz}(l_A, g, v_A) L_{s,low(RA)}(A=z) \right] \\
\times \quad & \sum_{y\in\Omega} \Bigg(P_{xy}(l_U, g, v_U) \\
\times \quad & \sum_{q\in\Omega} \left[P_{yq}(l_B, g, v_B) L_{s,low(UB)}(B=q) \right] \\
\times \quad & \sum_{v\in\Omega} \left[P_{yv}(l_C, g, v_C) L_{s,low(UC)}(C=v) \right] \Bigg) \Bigg] \Bigg]
\end{aligned}
\tag{1}
$$

where $P_{xy}(l_A, v_A)$ is the probability for base x to change into base y along a branch of length l_A, with rate category g from the Γ distribution and other evolutionary parameters v_A, $P(R=x)$ is the probability to have base x at the root R, and Ω is the set of possible states (for instance, $\Omega = \{A, T, C, G\}$ in case of a DNA alignment). $L_{s,low(RA)}(A=z)$ is the lower conditional likelihood of observing the data downstream from branch RA conditionally on the underlying

subtree and on having base z at node A. Note that computing the likelihood of a site when using a distribution over the evolutionary rates amounts to averaging the likelihoods of the site obtained when using each evolutionary rate in turn.

Computing the Likelihood with a Mixture Model on trees

As for the likelihood of a model where different rates are allowed, one can compute the likelihood of a model where one allows different trees. Consequently, to get the likelihood of a model whose parameters of interest are the trees that best describe the alignment, one can take at each site the average over the likelihoods obtained with each one of the trees that are considered.
This is summed-up in the following formula for the likelihood of a single site, where T represents the set of trees τ currently in use, and $|T|$ the number of trees in T:

$$L_{s,T} = \sum_{\tau \in T} \frac{1}{|T|} L_{s,\tau} \tag{2}$$

With such a formula, both rate heterogeneity and topology heterogeneity are taken into account, respectively by the gamma distribution and the Mixture Model on topologies. Once the likelihood of a Mixture Model over trees has been computed and maximized , it is possible to predict *a posteriori* the most likely tree for a given site (see below). This possibility can be used to uncover the recombination structure in an alignment.

Toy example: it is possible to optimize the topologies with a Mixture Model on trees

In a setting where we search for $|T|$ trees τ that describe an alignment, we try to find the set of $|T|$ trees whose likelihood as computed above in Eq. 2 is maximal. The object that is looked for is the set T itself. This can lead in principle to something different than simply using many times the maximum likelihood topology, as can be seen in this toy example, where $|T| = 4$, with 4 sites:

Topologies	Site 1 likelihood	Site 2 likelihood	Site 3 likelihood	Site 4 likelihood
Topology 1	10^{-2}	10^{-4}	10^{-4}	10^{-4}
Topology 2	10^{-4}	10^{-2}	10^{-3}	10^{-4}
Topology 3	10^{-4}	10^{-4}	10^{-2}	10^{-4}
Topology 4	10^{-4}	10^{-4}	10^{-4}	10^{-2}

In this example, the most likely topology is *Topology* 2, with a log-likelihood of $log(10^{-4} \times 10^{-2} \times 10^{-3} \times 10^{-4}) = -13$. However, if one were to use a Mixture Model on trees in which 4 trees are allowed, this should not simply result in the same *Topology* 2 topology being found in the 4 trees. Indeed, as for each site the average over the likelihoods for each topology is computed along the alignment, one obtains the following log-likelihood:

$$log(L_T) = log(\frac{10^{-2} + 3 \times 10^{-4}}{4} \times \frac{10^{-2} + 3 \times 10^{-4}}{4} \times \frac{10^{-2} + 2 \times 10^{-4} + 10^{-3}}{4} \times \frac{10^{-2} + 3 \times 10^{-4}}{4}) \approx -10.3$$

It is thus more likely on this example to use 4 different trees rather than a single tree. However, had the alignment been homogeneous, this model could have resulted in the same tree repeated 4 times, possibly with branch lengths differing between trees.

This example shows that in case of an alignment altered by a recombination event, a set of $|T|$ trees can be optimized to best account for the sequence evolution with a Mixture Model: it is not necessary that the tree topologies are specified before the search for the recombination breakpoint is undertaken.

A Phylogenetic Hidden Markov Model to detect recombination

The Mixture Model described above fails to account for an important property of the alignment: it is expected that the topology that best describes a given site has a high probability of properly describing the neighboring sites. Thus there is a dependency between sites, that can be modelled through the use of a Hidden Markov Model, whose hidden states are the topologies themselves. This model therefore belongs to the family of Phylo-HMMs. The rate heterogeneity is taken into account through a mixture model on rates, through the commonly used gamma distribution.

Computing the likelihood with the Phylo-HMM

The likelihood of the Phylo-HMM can be computed with the forward algorithm, as already explained in the phylogenetics framework by Felsenstein and Churchill [4]. We rapidly go through this algorithm here.

The algorithm starts from one end of the alignment and finishes at the other end; arbitrarily, we will start by the beginning of the alignment, at site 1, and end at site n. We suppose that individual site likelihoods have been already computed for all the trees. We note as $L_{1,\tau}$ the likelihood obtained with Felsenstein's pruning algorithm (no dependency between sites) at site 1 for the tree τ. The likelihood of the alignment up to site k with tree τ affected to site k is denoted $L_{\tau}^{(k)}$. The transition probability of going from tree τ at site k to tree τ' at site $k+1$ is written $P_{\tau,\tau'}$. We define as $|T|$ the total number of trees in the set T.

At the first site, the likelihood of the alignment up to site 1, given that site 1 has tree τ is simply the likelihood of tree τ for the site 1:

$$L_{\tau}^{(1)} = L_{1,\tau}$$

At the second site, the likelihood of the alignment up to site 2, given that site 2 has tree τ':

$$L_{\tau'}^{(2)} = L_{2,\tau'} \times \sum_{\tau \in T} P_{\tau,\tau'} L_{\tau}^{(1)}$$

This formula suggests a recursive scheme:

$$L_{\tau'}^{(k+1)} = L_{k+1,\tau'} \times \sum_{\tau \in T} P_{\tau,\tau'} L_{\tau}^{(k)}$$

The first part of the formula before the multiplication symbol is the classical likelihood of a tree for site k, which can be obtained through Felsenstein's pruning algorithm [14] as in equation 1. The dependency between sites is introduced through the second part of the formula. At the end of the alignment, at site n, the total likelihood of the alignment given the set of trees T is computed as follows:

$$L_{Phylo-HMM} = \sum_{\tau \in T} \frac{1}{|T|} \times L_{\tau}^{(n)} \tag{3}$$

In our model, the transition probability of going from tree τ at site k to tree τ' at site $k+1$, $P_{\tau,\tau'}$ is defined as follows, with the help of the autocorrelation parameter λ:

$$P_{\tau,\tau'} = \lambda \delta_{\tau,\tau'} + \frac{1-\lambda}{|T|}$$

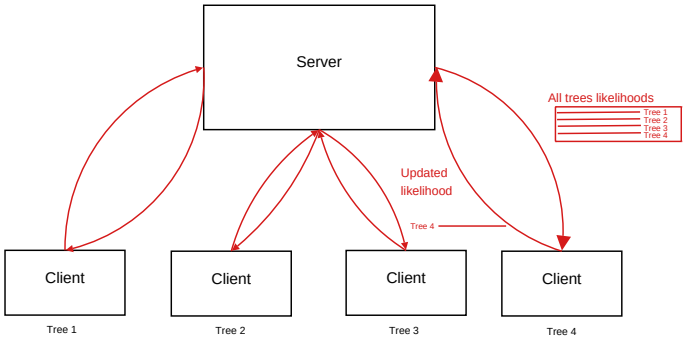
Here, $\delta_{\tau,\tau'}$ is the Kronecker delta function, which is 1 when $(\tau = \tau')$ and 0 otherwise. This means that at any site, there is a constant probability λ that the same tree is kept for the next site, and a probability $1-\lambda$ that another tree is drawn for the next site, with the possibility that the same tree is drawn again.

Since one can compute the likelihood of the alignment with the Phylo-HMM, all parameters can be estimated in the maximum likelihood framework (or in a bayesian framework). Therefore in our program, both the trees (topologies, branch lengths, parameters of the models) and the parameter λ are estimated by optimizing the likelihood as computed in equation 3, through the same algorithm as PhyML for common parameters, and through Brent's numerical optimization algorithm [15] for the autocorrelation parameter λ.

Exploring the Space of Tree Topologies with a Mixture Model on trees or with a Phylogenetic Hidden Markov Model

The problem of optimizing $|T|$ trees simultaneously is different from the problem of optimizing a single topology $|T|$ times. At any given time, a topology is to be optimized *taking into account the other topologies*. Indeed, if each topology were optimized independently of the other topologies, the result would be $|T|$ identical trees: this would have been equivalent to solving the single tree optimization problem $|T|$ times, in parallel.
A parallel algorithm based on a server-client architecture, as described in Fig. 2, allows to acknowledge the dependencies between topologies.

Figure 2: Server-Client architecture to efficiently find a set of topologies that best describe the alignment.



The server exchanges data with clients. For each set of communications between the server and a client, one red arrow corresponds to the sending by the server to the client of a matrix containing all the site likelihoods for all the topologies, and the other one corresponds to the sending by the client to the server of an optimized likelihood vector.

In this algorithm, each client is affected a topology, that it tries to refine through commonly used tree search algorithms. However, while in common algorithms such as PhyML the client would simply try to maximize the likelihood of the topology, here it needs to maximize the likelihood of the MM or of the phylo-HMM as a whole, by only modifying the topology it has been affected, while taking into account the other topologies. For instance, in the Mixture Model, the likelihood function each client tries to maximize thus is $L_{Tree\ mixture} = \prod_s \sum_{\tau \in T} \frac{1}{|T|} L_{s,\tau}$, which implies that each client needs vectors of site likelihoods obtained from the other clients. The dependency between topologies is only taken into account through a shared matrix of likelihood vectors.

The algorithm has been summed up in the pseudo-code below.

Algorithm 1 Searching for the most likely set of trees T.

```
likelihood_threshold=1e-6
MAXIMUM=1e6
|T|=2
if (server) {
    get alignment aln
    set_of_trees T = Generate(|T|,aln)
    Create |T| clients
    send_all alignment
    send trees T
    likelihood_matrix = receive_all_likelihood_vectors()
    oldlk=compute_likelihood(likelihood_matrix)
    send_all(likelihood_matrix)
    diff=MAXIMUM
    while (diff>likelihood_threshold) {
        receive(likelihood_vector)
        update(likelihood_matrix)
        newlk=compute_likelihood(likelihood_matrix)
        diff=newlk - oldlk
        oldlk=newlk
        send_all(likelihood_matrix)
    }
    send_all(stop_signal)
    output_server_results
}
else if client {
    receive alignment
    receive tree
    compute_likelihood
    send(likelihood_vector)
    receive(likelihood_matrix)
    while (not stop_signal)
    {
        optimize(tree, likelihood_matrix)
        send(likelihood_vector)
    }
    output_client_results
}
```

At the beginning of the program, the number of topologies to consider needs to be set, as this algorithm is not able to estimate the appropriate number of trees $|T|$ to consider to describe the history of an alignment; in the pseudo-code above, it has been set to 2. In practice, setting this parameter should hardly be a problem, as a gene sequence should not harbour more than two (detectable) different evolutionary histories; it is however possible to specify more than two topologies to be searched for in a single alignment. At the begining of the algorithm, the function "Generate" divides the alignment in $|T|$ equal parts and builds a BIONJ [16] tree for each part. This results in $|T|$ trees used as starting topologies for the bulk of the algorithm (alternatively, the user can also provide $|T|$ starting trees). Each client then receives the alignment and a tree it is in charge of, computes the likelihood of this topology, and returns a vector of site likelihoods to the server. The server assembles all vectors into a matrix, that is sent to all clients. Each client subsequently modifies the specific tree it is in charge of, in order to maximize, $L_{Tree\ mixture}$ or $L_{Phylo-HMM}$. Periodically, it sends an updated vector of site likelihoods to the server, which updates the likelihood matrix containing all likelihood vectors. This updated matrix is subse-

quently sent to all clients, so that they continue optimizing their topologies acknowledging the most recent changes in other topologies. In practice, communications between the server and the client are asynchronous, so that slowly-computing clients do not slow down the other clients. For the Phylo-HMM, the auto-correlation parameter λ is also exchanged between the server and the clients, and optimized by the server every ten times it receives a likelihood vector from one of its clients.

This algorithm has been implemented to function with both the MM and with the Phylo-HMM (where the autocorrelation parameter λ is exchanged between the server and clients, and periodically optimized by the server) in the PhyML-Multi program, based on PhyML v.2.4.4 code [15]. This program can take advantage of a multi-processor or multi-core machine, by dispatching clients in charge of trees to different processors. It has been compiled and tested on Linux machines and is available on request.

As a result, each client outputs an optimized topology, and the server outputs the matrix containing site likelihoods computed with each topology. If there have been recombination events in the history of the alignment, there should be stretches of sites whose most likely topology is the same. Through segmenting the matrix of site likelihoods, one should be able to uncover these stretches of sites with a common history. The Phylo-HMM can directly output a most likely segmentation; on the other hand, the Mixture Model does not provide such a segmentation.

Segmenting the matrix of site likelihoods output by the Mixture Model

Methods to partition an alignment

Common approaches to segmentation involve the use of sliding windows, Hidden Markov Models or of the Maximum Predictive Partitioning algorithm (MPP algorithm, [17, 18]). We have chosen not to use sliding windows, as the fixed size of the sliding window does not allow to precisely pinpoint the recombination events. Both the MPP algorithm and the HMM approach rely on a statistical approach to segment a sequence: given a set of models, they infer the most likely partitioning of the sequence into these models. In our case, the models are the trees themselves, and the sequence is the alignment. For each model, the site likelihoods have been previously computed by the MM. The partitioning of the alignment therefore is done according to these site likelihoods, which are used during the computation or the segmentation likelihood.

The HMM approach permits to directly estimate a partitioning, which depends upon the transition probabilities between models. These transition probabilities can be estimated with the Baum-Welch algorithm. However, they constrain the length of the stretches of sites that share the same model to follow a geometric distribution. In the case of the detection of recombination, this can be problematic because there is no reason that the lengthes of all segments sharing a unique history should follow such a distribution.

The MPP algorithm on the other hand does not require that transition probabilities are set, and thus does not constrain the sizes of the segments. However, as a consequence, the MPP algorithm does not provide a single most likely partitioning, but outputs a most likely partitioning in two segments, three segments, four segments... In the end the user is faced with a range of most likely partitionings, among which a choice is to be made according to some criterium.

Estimating the number of segments with the MPP algorithm

As the number of segments increases, the likelihood of the segmentation generally also increases, not necessarily because adding a segment reveals a significant property of the alignment, but also because adding a segment may permit to better fit a non-significant heterogeneity in a particular part of the alignment. In other words, the improvement in likelihood observed when the number of segments increases is due to the fitting of the "noisy" part of the signal rather than the meaningful part.

Such non-significant gains in likelihoods can also be seen in alignments where sites have been randomly swapped, erasing the meaningful signal of the recombination structure, but where non-significant heterogeneities are expected to be found simply by chance. Therefore the comparison between the true alignment and randomized versions of the alignment permits to distinguish improvements in the likelihood of a partitioning due to the uncovering of a homogeneous segment coming from a past recombination event from "noise" improvements in the likelihood, due to the fitting of non-significant heterogeneities.

To get an estimate of the number of segments in an alignment, the following protocol is thus applied, for each number i of segments in $[1; n]$, with n defined *a priori* by the user:

- the likelihood of the most likely partitioning in i segments is computed using the MPP algorithm, and stored in the value L
- the matrix of site likelihoods is randomized 100 times by swapping columns of site likelihoods (which is equivalent to swapping sites in the alignment), and for each of these 100 replicates, the likelihood of the most likely partitioning is computed using the MPP algorithm; the average of these 100 likelihood replicates is computed and stored in the value $\bar{l}$
- the value $L^* = \frac{L}{\bar{l}}$ is computed and used as a normalized likelihood for the partition in i segments

In the end, all normalized likelihoods can be compared; the partitioning with the highest normalized likelihood is considered as the most reasonnable partitioning.

Tests of the Mixture Model and the Phylo-HMM Model

HMM segment length follows a geometric law of parameter λ, the autocorrelation parameter. This law might not be appropriate to model the length segments in an alignment where there has been recombinations. The MPP approach does not introduce such a constraint on segment length, and may therefore produce different results from the HMM segmentation. The Phylo-HMM approach and the MM + MPP approach may therefore complement each other, each having defaults that the other does not have. This suggests that both approaches should be used in parallel, and their results compared. In this purpose, we used simulations.

Simulation procedure

The first 100 trees from the PhyML test set [15] were selected. These trees contain 40 leaves, were designed to resemble real-life datasets and should therefore provide an appropriate test-set. An alignment affected by a recombination is an alignment whose part is best described by a particular tree, and part by another tree. In the most difficult instances, the two trees corresponding to the two parts of the alignment differ by a single clade whose position is in one place in the first tree, and another place in the other tree. To obtain such pairs of trees, each of the 100 trees was subjected to a Subtree Prune and Regraft operation (SPR), in which a subtree is detached from the tree and attached in another position. This yielded pairs of trees separated by one recombination event, with Robinson and Foulds distances ranging from 2, when the SPR regrafted the pruned subtree very close to its original position, to 30, when the pruned subtree was regrafted far from its original position. Alignments harbouring a recombination event were simulated by evolving a portion of an alignment according to one of the 100 trees and the rest of the alignment according to the same tree modified by the SPR. For each pair of trees, 9 1000-nucleotide alignments were simulated with k sites according to one tree and $1000 - k$ sites according to the other tree, with k taking the values $100, 200, 300, 400, 500, 600, 700, 800, 900$. Seq-Gen [19] was used to simulate sequences, with the GTR model [20] and a continuous gamma rates across site distribution with parameter alpha set to 0.8.

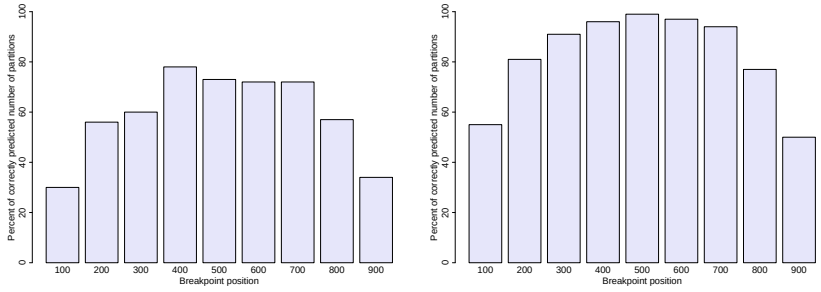
Reconstruction of the recombination structure with the Mixture Model and the Hidden Markov Model

Both the Mixture Model and the Hidden Markov Model were applied to the simulated datasets. The number of trees were set to two for both examples, as none of the programs is able to estimate the right number of trees to consider to faithfully describe an alignment. The evolutionary model used was HKY [21] with a gamma distribution discretized in four classes to account for accross site rate variation. The reconstruction model therefore does not exactly correspond to the simulation model, as would be the case in a realistic setting where sequences have evolved according to an unknown and complex process.

Ability to detect the right number of segments

The reconstruction models should detect two parts in the alignment. Figure 3 shows that both models have a recovery rate that is dependent upon the position of the breakpoint. If the breakpoint is too close to the begining or the end of the alignment, the recovery rate is lower than if the breakpoint is more central. This is likely because lengths such as 100 or 200 nucleotide sites contain too little information to properly reconstruct a tree topology. Such values may therefore represent the statistical limit below which our models cannot detect recombination. The Phylo-HMM is superior to the MM in all cases, which indicates that acknowledging that it is highly probable that neighbor sites have the same most likely tree improves the breakpoint detection.

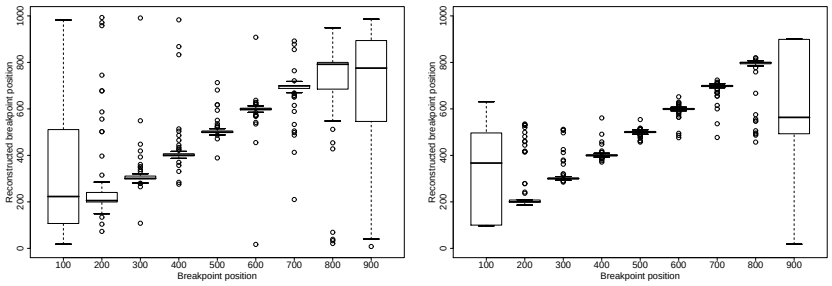
Figure 3: Ability of the Mixture Model (left) and Phylo-HMM (right) to detect the number of segments in simulated alignments.



Ability to detect the breakpoint position

Both the MM and the Phylo-HMM most often detect two segments in the alignment. In such cases, Figure 4 shows that the precision with which the breakpoint is predicted displays the same dependency upon the length of the smaller segment as the ability of the models to detect the number of segments. The phylo-HMM seems slightly better than the MM in detecting the precise breakpoint position when the smallest partition is ≥ 200 bases long. However, the Phylo-HMM is less good that the Mixture Model when the smallest partition is 100 bases long. This is likely a manifestation of the bias introduced by the geometrical distribution of segment length in the HMM.

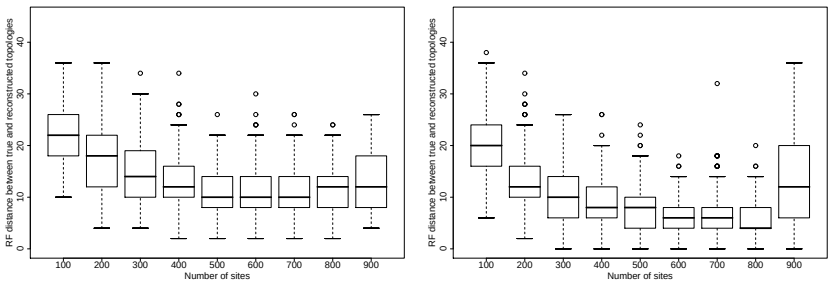
Figure 4: Ability of the Mixture Model (left) and Phylo-HMM (right) to detect the breakpoint position in simulated alignments.



Ability to recover the true topologies

On average, the Phylo-HMM is better at recovering the trees used in the simulation than the MM, and both models find it easier to get good trees if the alignment that has been simulated along them is long. However, the quality of the reconstructed trees finds an optimum for alignments that are 600 to 800 sites, not longer. When one of the two topologies found in the alignment represents only 100 sites, both topologies, the one found in 100 sites and the one found in 900 sites, are less well reconstructed.

Figure 5: Ability of the Mixture Model (left) and Phylo-HMM (right) to recover topologies from simulated alignments.



Computation times

Computations were run on the IN2P3 computing centre, on computers ranging from 2.2 to 2.8 GHz. It took on average 9min48s for the Mixture Model implementation to give a result on the simulations, while only 3min45s for the phylo-HMM. The additional optimization of the autocorrelation parameter has not resulted in an increased computational time, but a decrease, perhaps because the HMM ensures that the set of sites pleading for a given topology is more stable throughout the tree space search than when the MM is used. However, both models are very efficient on datasets containing 40 sequences and on single desktop computers.

Conclusions on the simulations

Overall, the Phylo-HMM is better able to uncover the recombination structure of simulated alignments, since it more often finds the right number of segments, is more accurate at pinpointing the recombination breakpoint, and also recovers trees closer to the true trees. This is probably because the HMM takes into account the dependency of neighboring sites. However, for the smallest segments (100 site long), the Phylo-HMM appears less good than the MM at predicting the breakpoint position, probably because a single autocorrelation parameter is used to describe

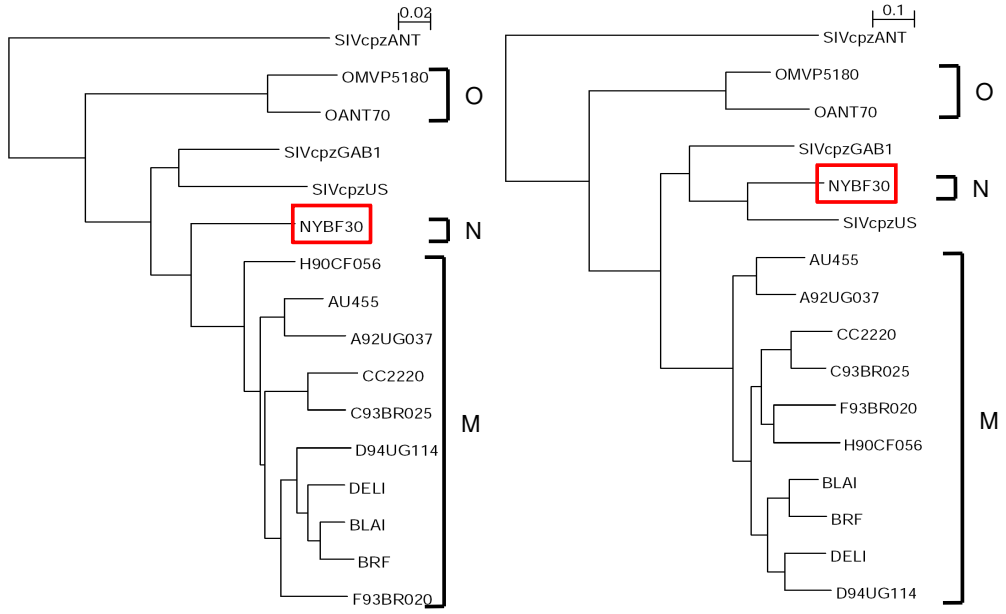
the length of both segments, the one that is 100 bases long, and the one that is 900 bases long. Allowing different autocorrelation parameters for different hidden states (here phylogenetic trees) might correct this weakness; however, it would also increase the number of parameters of the model. Instead, we recommend using both the MM and the Phylo-HMM to analyse datasets, as the advantages of one compensates the drawbacks of the other.

Application to real protein sequences

Several studies have unveiled recombination events in viruses, for instance in HIV viruses. In 1999, Gao *et al.* discovered that a recombination event in a chimpanzee host was at the origin of the YBF30 (group N) HIV-1 virus: the begining of the genome of YBF30 was most closely related to group M whereas the rest of its genome was most closely related to a chimpanzee virus, SIVcpzUS. They based this conclusion on first a sliding window analysis where divergence between pairs of sequences was computed, and second the reconstruction of trees for two portions of the alignment, on each side of a putative recombination breakpoint, which had been identified by eye. Likelihood tests confirmed the recombination event, showing that the first part of the alignment rejected the tree obtained for the second part, and *vice-versa*.

This study therefore provides a good test of the ability of the Mixture Model and the Phylo-HMM to detect recombination in natural conditions. The two models were run on the alignment from Gao *et al.*, setting the number of trees to two. The Mixture Model predicted two breakpoints, one at position 95, and the other at position 1354. The phylo-HMM predicted only one breakpoint, at position 1353. The two models therefore agree on the presence of a breakpoint around position 1353, which falls very close to the recombination breakpoint determined by eye in the original analysis, at position 1400. The additional breakpoint predicted by the MM is more uncertain as it is not detected by the Phylo-HMM: it might be due to the higher sensitivity of the MM when one of the two segments is small, or might be non-significant. Interestingly, both the MM and the Phylo-HMM uncover the shifting position of YBF30, which first is close to group M sequences, and then close to SIVcpzUS (see figure 6 for trees found with the Phylo-HMM).

Figure 6: Trees found by the Phylo-HMM on Gao *et al.* data. The trees found by the Mixture model are nearly identical.



This example shows that the Phylo-HMM is also efficient on real sequence datasets. The use of such a program offers an improvement over the sliding-window approach taken by Gao et al. : indeed, if one is to look for a recombination event in any sequence, all sequences are to be analysed two by two, which, for the 16 sequences present in the tree amounts to looking at $16 * 15/2 = 240$ plots of divergence. With programs such as ours, only two steps are required, as advocated by Chan *et al.* [22]: first a statistical measure to detect the occurence of recombination needs to be applied; if positive, our programs can then be used to precisely pinpoint the recombination breakpoint and reconstruct phylogenetic trees. This way, all the sequences are analysed at once, and the user input is minimal. Eventually, statistical tests such as implemented in Consel [23] can be applied to confirm the occurence of recombination.

Conclusion

In this article, a Mixture Model and a Phylogenetic Hidden Markov Model to detect recombination were presented. Both methods were tested on synthetic datasets, which showed that the Phylo-HMM was superior to the Mixture Model in most circumstances, except when the recombination event had only affected a small portion of the alignment. Notably, both methods were highly efficient. The analysis of an already published dataset showed that the models could successfully uncover recombination breakpoints and topologies. Future improvements might include searching for the appropriate number of topologies to use, or constraining the topologies on each side of a breakpoint to differ by no more than one rearrangement.

References

[1] Posada D: **Evaluation of methods for detecting recombination from DNA sequences: empirical data.** *Mol Biol Evol* 2002, **19**(5):708–717.

[2] Bruen TC, Philippe H, Bryant D: **A simple and robust statistical test for detecting the presence of recombination.** *Genetics* 2006, **172**(4):2665–2681, [[http://dx.doi.org/10.1534/genetics.105.048975]].

[3] McGuire G, Wright F, Prentice MJ: **A Bayesian model for detecting past recombination events in DNA multiple alignments.** *J Comput Biol* 2000, **7**(1-2):159–170, [[http://dx.doi.org/10.1089/10665270050081432]].

[4] Felsenstein J, Churchill GA: **A Hidden Markov Model approach to variation among sites in rate of evolution.** *Mol Biol Evol* 1996, **13**:93–104.

[5] Husmeier D: **Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models.** *Bioinformatics* 2005, **21 Suppl 2**:ii166–ii172, [[http://dx.doi.org/10.1093/bioinformatics/bti1127]].

[6] Husmeier D, Wright F, Milne I: **Detecting interspecific recombination with a pruned probabilistic divergence measure.** *Bioinformatics* 2005, **21**(9):1797–1806, [[http://dx.doi.org/10.1093/bioinformatics/bti151]].

[7] Kedzierska A, Husmeier D: **A heuristic Bayesian method for segmenting DNA sequence alignments and detecting evidence for recombination and gene conversion.** *Stat Appl Genet Mol Biol* 2006, **5**:Article27, [[http://dx.doi.org/10.2202/1544-6115.1238]].

[8] Suchard MA, Weiss RE, Dorman KS, Sinsheimer JS: **Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage.** *Syst Biol* 2002, **51**(5):715–728, [[http://dx.doi.org/10.1080/10635150290102384]].

[9] Minin VN, Dorman KS, Fang F, Suchard MA: **Dual multiple change-point model leads to more accurate recombination detection.** *Bioinformatics* 2005, **21**(13):3034–3042, [[http://dx.doi.org/10.1093/bioinformatics/bti459]].

[10] Minin VN, Dorman KS, Fang F, Suchard MA: **Phylogenetic mapping of recombination hotspots in human immunodeficiency virus via spatially smoothed change-point processes.** *Genetics* 2007, **175**(4):1773–1785, [[http://dx.doi.org/10.1534/genetics.106.066258]].

[11] Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW: **GARD: a genetic algorithm for recombination detection.** *Bioinformatics* 2006, **22**(24):3096–3098, [[http://dx.doi.org/10.1093/bioinformatics/btl474]].

[12] Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW: **Automated phylogenetic detection of recombination using a genetic algorithm.** *Mol Biol Evol* 2006, **23**(10):1891–1901, [[http://dx.doi.org/10.1093/molbev/msl051]].

[13] Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**(4):406–425.

[14] Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17**(6):368–376.

[15] Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**(5):696–704.

[16] Gascuel O: **BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.** *Mol Biol Evol* 1997, **14**(7):685–695.

[17] Gueguen L: **Segmentation by maximal predictive partitioning according to composition biases.** In *Computational Biology, Volume 2066 of* LNCS. Edited by O G, M S, Springer-Verlag 2001:32–45.

[18] Gueguen L: **Sarment: Python modules for HMM analysis and partitioning of sequences.** *Bioinformatics* 2005, **21**(16):3427–3428, [[http://dx.doi.org/10.1093/bioinformatics/bti533]].

[19] Rambaut A, Grassly NC: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Comput Appl Biosci* 1997, **13**(3):235–238.

[20] Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates**. *J. Mol. Evol.* 1984, **20**:86–93.

[21] Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22**(2):160–174.

[22] Chan CX, Beiko RG, Ragan MA: **Detecting recombination in evolving nucleotide sequences.** *BMC Bioinformatics* 2006, **7**:412, [[http://dx.doi.org/10.1186/1471-2105-7-412]].

[23] Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Syst. Biol.* 2002, **51**(3):492–508.

Acknowledgements

This work was supported by Agence Nationale de la Recherche (GIP ANR JC05_49162) and by the Centre National de la Recherche Scientifique.

9

Simultaneous Inference of a Species Tree and of Gene Trees

Figure 9.1: *Gene trees are deformed shadows of the species tree. If one wants to infer a species tree, the best way to do so is to use models of gene family evolution. Painting by Françoise Boussau-Janon.*

Article 3 is a clear illustration that gene trees can differ from species trees. In this article, incongruences were thought to be due to gene transfer, but without any proof. In fact, other biological phenomenons can render gene trees different from species trees (for more on this, see article 10).

In the present article, I built a model to try and infer a species tree when gene trees may differ from it because of gene duplications and gene losses. This model was implemented in a program that can run on several computers simultaneously, and that could be easily modified to cope with other causes of gene tree/species tree incongruences.

This article has not been submitted yet.

Simultaneous inference of gene trees and species tree in the presence of duplications and losses

Bastien Boussau*, Eric Tannier, Manolo Gouy, Vincent Daubin

September 10, 2008

*Corresponding author

Université de Lyon ; Université Lyon 1 ; CNRS ; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France.

E-mail: boussau@biomserv.univ-lyon1.fr

Abstract

Species trees are usually built as an average of the signal of several genes. However, several biological processes can affect gene families to the extent that gene trees may strongly differ from the true species tree. Duplications and losses are two such processes. In order to reconstruct a species tree from genes, we propose to model gene family evolution in the presence of gene duplication and loss, and consequently separately infer gene trees and species tree. In this model, each branch of the species tree is associated to particular duplication and loss probabilities. We explain how one can compute the likelihood of a species tree with such a model, what algorithms can be used with it, and present a natural parallel architecture to speed-up the computations. In addition to duplication and loss, this framework could be easily extended to use models of gene transfers or of trans-specific polymorphism.

Introduction

When inferring speciations that occured millions or billions of years ago, one is tempted to use as much data as possible. Since 1964 and the first molecular species tree built from seven sequences containing nineteen amino-acids each (Doolittle and Blombaeck, 1964), this tendency has been facilitated by progresses in sequencing techniques and computer science. Recently, data from whole genome or large scale EST sequencing projects have permitted to analyse dozens of genes for dozens of species simultaneously (Ciccarelli *et al.*, 2006; Delsuc *et al.*, 2006; Dunn *et al.*, 2008). However, the resulting trees are not perfectly resolved and still contain weakly supported bipartitions.

These uncertainties may result from closely spaced cladogenesis events (Degnan and Rosenberg, 2006), model misspecification (Felsenstein, 1978; Weisburg *et al.*, 1989), or the way orthology has been defined. Indeed, both currently available methods, the concatenation and the supertree approaches (Delsuc *et al.*, 2005), require that each species whose history is to be reconstructed is represented in each alignment by no more than one gene. This amounts to betting that a gene present in single copy in all genomes under consideration is a *bona fide* representative of the species phylogeny, and has not undergone series of duplications and losses that might complicate its history. However, it is not unfrequent that gene families underwent duplications, so that a gene family history may not simply be the mirror image of the speciation history, to the point that one species may harbour several paralogous copies of the same gene. In such cases of obvious paralogy, if not too many species possess several genes from a single gene family, before phylogenetic reconstruction, the phylogeneticist chooses one copy based on *a priori* knowledge of some phylogenetic relationships or based on the consideration of similarity scores, or altogether discards all families which contain paralogous genes. Such an additional step conditions the analysis, decreases the amount of data submitted to phylogenetic reconstruction, and is highly dependent upon subjective choices of the researcher.

Here we present a hierarchical probabilistic model of gene tree reconciliation and sequence evolution. It provides a robust and comprehensive approach to species phylogeny, able to analyse thousands of gene families, paralogs included, and simultaneously reconstruct highly resolved species and gene family trees. Additionally, it also supplies reconciliated gene trees, and decorates the branches of the species tree with counts of gene duplication and gene loss events. Because we have not been able to produce a fully working algorithm in time, no example of application will be provided.

A hierarchical model of gene family evolution

In Eukaryotes, gene families evolve mainly through duplication, loss and sequence divergence. The probabilistic modelling of sequence divergence has been the object of a large body of litterature, starting with the model of Jukes and Cantor (Jukes T.H., 1969), and continuously improving with the inclusion of variablility of rates of evolution among sites (Yang, 1994; Felsenstein and Churchill, 1996), and among branches (Tuffley and Steel, 1998; Galtier, 2001), variability of models of evolution among sites (Pagel and Meade, 2004; Lartillot and Philippe, 2004), and among branches (Yang and Roberts, 1995;

Galtier and Gouy, 1998; Foster, 2004; Gowri-Shankar and Rattray, 2006; Blanquart and Lartillot, 2006; Boussau and Gouy, 2006; Gowri-Shankar and Rattray, 2007; Blanquart and Lartillot, 2008), to cite only a few recent examples. The statistical modelling of duplication and loss is more recent. Parsimony reconstructions of gene family evolution were first developped in 1979 (Goodman *et al.*, 1979), and since then have been the object of several articles attempting to improve the algorithms (Mirkin *et al.*, 1995; Guigo *et al.*, 1996; Page and Charleston, 1997; Zmasek and Eddy, 2001; Bansal and Eulenstein, 2008; Wehe *et al.*, 2008). More recently, statistical models of gene family evolution have been developped (Arvestad *et al.*, 2003; Dubb, 2005), in which gene duplications and gene losses are modelled by a birth-death process, with birth and death probabilities shared by all lineages. As for models of sequence evolution that are able to infer "hidden" substitutions, and contrary to methods based on parsimony, the use of a birth-death process permits to infer events of gene duplications and gene losses that have not left any trace on the resulting topology. This biological realism however does not come without a cost, and it seems difficult for a program implementing such a model to analyse thousands of genes simultaneously, for dozens of species, to infer a species tree. Such datasets however are already available, as more than 83 whole genomes from Eukaryotes have been sequenced and published to date (Liolios *et al.*, 2008). A model that could achieve the analysis of such datasets without forfeiting too much on the realism side is therefore needed.

A simplification inherent to the models of Arvestad *et al.* (2003) and Dubb (2005) is that duplication and loss probabilities are constant over the whole species phylogeny. However, all branches in a species phylogeny have not undergone the same amount of duplications and losses: modelling such events with a single probability for duplications and another one for losses, independently of the position of the event in the species tree, may not be appropriate.

We therefore choose to associate a particular pair of duplication and loss rates $\{d_i, l_i\}$ to each branch i of the species tree. To compute the likelihood of a rooted gene family tree, we use a reconciliation algorithm of Zmasek and Eddy (2001) (Fig. 1), in which nodes of the gene tree are mapped onto nodes of the species tree : the basic principle is to map the nodes of the gene tree to the nodes of the species tree according to the following principle. For a node u, $L(u)$ denotes the set of species that have a gene that is a descendant of u, and for a set of species S, $lca(S)$ denotes the node that is the last common ancestor of all species in S. Then a node u in the gene tree is mapped to $\lambda(u) = lca(L(u))$. Moreover, "hidden nodes" in the gene tree, *i.e.* nodes that are not visible in the gene tree because one of their two descendants has been lost, are also mapped to nodes of the species tree.

In Zmasek and Eddy (2001), this mapping aims at providing a most parsimonious scenario of duplications and losses in the gene family that explains the difference between a rooted gene tree and a rooted species tree: a duplication event is associated to a node u of the gene tree if it has as least one child v such that $\lambda(u) = \lambda(v)$, and a loss event is infered every time v is a child of u in the gene tree, while $\lambda(u)$ is not a child of $\lambda(v)$ in the species tree.

Here, we do not infer scenarios, but integrate on all scenarios that may explain this difference, to compute a likelihood. So for example if a gene loss is infered by the algorithm of Zmasek and Eddy (2001), we consider the possibility

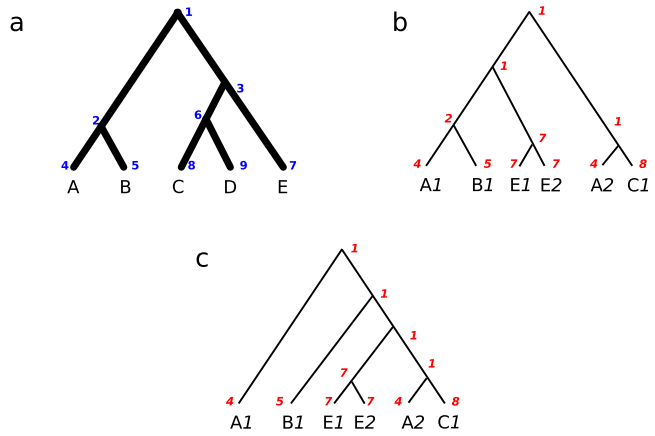


Figure 1: Mapping between the species tree and a gene tree. a: Species tree, with its nodes numbered, from 1 at the root to higher numbers as nodes are further from the root. b: Gene tree rooted at a node showing the most parsimonious duplication/loss scenario along the tree. The nodes are numbered in reference to the species tree. c: Same gene tree rooted in a random position, numbered in reference to the species tree. The most parsimonious rooting position is on a branch linked by any extremity to a node numbered as 1.

of a duplication followed by two losses, or even more unparsimonious scenarios. This integration thus accounts for hidden events, just as models of sequence evolution account for hidden substitution events.

To each node u of the species tree, let $\mathcal{T}(u)$ be the subforest of the gene tree which is the graph induced by all nodes mapped to u by the function λ. For a component T of $\mathcal{T}(u)$, a vertex is an *exemplar* if it is a leaf, or an internal node of degree 2, or the root if it has degree 1. Every component T corresponds to a number of duplications in the most parsimonious solution of Zmasek and Eddy (2001), which is equal to the number of exemplars minus one. In particular, if T is composed of a single node, it corresponds to an evolution without duplication event in the branch i of the species tree leading to u. The number of exemplars of a component T of $\mathcal{T}(u)$ gives the number of paralogous genes obtained by the duplication process (one if T is a single node).

To take into account the possibility of unparsimonious scenarios, we use a birth-death process, where birth corresponds to duplication, and death corresponds to loss. Not all possible scenarios are yet taken into account, as we take for granted the number of paralogous genes after a duplication process witnessed by all the exemplars of a component of $\mathcal{T}(u)$. In case we count 0 exemplar in a component, we neglect the probability that more paralogous copies have been generated at node u, and lost later. We believe this approximation may not be too harmful to our model.

Formulas for computing the probability $P_u(k)$ that a component T of $\mathcal{T}(u)$ has k exemplars, or that a loss of gene is inferred at node u (this corresponds to the case $k = 0$) can be found in Thorne *et al.* (1991) who used a birth-death

process to model insertions and deletions in sequences:

$$\begin{aligned} P_u(0) &= l_u \times \beta \\ P_u(k) &= (1 - d_u \times \beta) \times (1 - l_u \times \beta) \times (d_u \times \beta)^{k-1} \text{ with } k \in [1; +\infty[\end{aligned}$$

where

$$\beta = \frac{1 - e^{d_u - l_u}}{l_u - d_u \times e^{d_u - l_u}} \tag{1}$$

Using these branch probabilities, and assuming the evolution of a gene along a branch is independent of its evolution along another branch, one can compute a likelihood for a reconciliation (the duplication/loss evolution of a gene family), $L(reconciliation)$, as follows (Equation 2):

$$L(reconciliation) = \prod P_u(k_u) \text{ with } k_u \in [0; +\infty[\tag{2}$$

This product is computed over all nodes u of the species tree, and for each node u, over all components of $\mathcal{T}(u)$ and all losses events inferred by the algorithm of Zmasek and Eddy (2001). Then k_u is the number of exemplars in a component T of $\mathcal{T}(u)$, or 0 for a gene loss.

Equation 2 allows to compute the likelihood of a reconciliation of losses and duplications, but does not take into account sequence evolution. A *family* likelihood, which combines the likelihood of a reconciliation and the likelihood *sensu* Felsenstein (Felsenstein, 1981) of a gene tree, is obtained as follows:

$$L(family) = L(reconciliation) \times L_{Felsenstein}(gene\ tree) \tag{3}$$

$L(family)$ permits to compute the likelihood of a gene family given a species tree. This likelihood can then be maximized with respect to the gene tree, or with respect to a species tree. If many gene families are to be analysed in parallel, assuming that they evolved independently from each other, both gene trees and species tree can be optimized, by maximizing the following likelihood (eq. 4):

$$L(Species\ \&\ gene\ trees) = \prod_{G \in \{All\ gene\ families\}} L_G(family) \tag{4}$$

Such a joint search requires specific algorithms.

Algorithms to simultaneously infer species and gene trees

Finding the best reconciliation between a gene tree and a species tree: rooting the gene tree

For a given rooted binary gene tree and a given rooted binary species tree, finding the reconciliation mapping λ is achieved through an algorithm akin to Zmasek and Eddy (2001). However, gene trees are not naturally rooted, unless they are inferred through molecular clock models (Zuckerkandl and Pauling, 1962; Kumar, 2005) or through non-reversible models of evolution (Yang and

Roberts, 1995; Galtier and Gouy, 1998; Huelsenbeck *et al.*, 2002; Yap and Speed, 2005; Boussau and Gouy, 2006). And even in such cases, there may be little signal for the position of the root. Our procedure therefore roots gene trees and searches for the root position on the best species tree. For a given rooted species tree, the best root position on the gene tree can be found by computing the reconciliation likelihood for all possible root positions, and then taking the root position that maximizes the likelihood. With this procedure, the most likely root may differ from the most parsimonious root. However, this approach is also very time-consuming, as for each gene tree with n leaves and each species tree with n_s leaves, $2n-3$ root positions need to be tried. In the algorithm of Zmasek and Eddy (2001), a mapping between nodes of the gene tree and nodes of the species tree must be obtained, through a gene tree traversal, which has a complexity of $O(n)$ in most cases. So the simplest algorithm to compute all reconciliations for the $2n-3$ possible root positions would imply $2n-3$ tree traversals, with a complexity in $O(n^2)$. Then scores need to be computed for each of the $2n-3$ reconciliations. Computing the score of a reconciliation uses equation 2, which involves $O(n_s)$ operations. If we consider that $n_s \approx n$, we obtain a total complexity on the order of $O(n^3)$ to get likelihood scores for all possible rootings, which permits to choose the most likely one. To minimize the number of tree traversals, Chen *et al.* (2000) relied on a double-recursive tree traversal algorithm, which overall permitted to compute all scores associated with the $2n-3$ reconciliations in $O(n^2)$. Instead, we used another approach avoiding to try all root positions by considering that the most likely one should not be too distant from the most parsimonious one: we save time by not computing all $2n-3$ root scores. The *rationale* is based on the first step of Zmasek and Eddy (2001) algorithm, which maps species tree nodes onto gene tree nodes (Fig. 1). If the species tree has been numbered so that nodes close to the root have smaller numbers than nodes far from the root, given an arbitrary root position chosen on the gene tree, the most parsimonious rooting (Fig. 1b) will be on a branch linked to a node whose index is the smallest on the gene tree (Fig. 1c) (proof not shown).

Therefore, once a mapping has been computed given an arbitrary rooting, to get the most parsimonious rooting, only branches linked to nodes with the smallest index need to be tried. In our model, the most parsimonious rooting may not be the most likely one; however, we assume that the most likely rooting will not be very distant from the most parsimonious one. Consequently, to find the root of a gene tree, the following procedure is applied:

1. Choose an arbitrary root and compute the mapping
2. Compute reconciliation scores obtained when rooting on a branch linked to a node whose index is inferior to a certain *threshold*
3. Choose the most likely rooting

The *threshold* used is user-defined through a parameter t. For a gene tree with smallest node index s, $threshold = s + t$.

Finding the best gene tree given a species tree: optimizing the gene family likelihood

The preceding section explained how particular gene and species trees could be compared to compute the most likely reconciliation score by looking for the gene tree root. If gene trees could be known *a priori*, this reconciliation score would be enough to search for the species tree by maximizing the product of all reconciliation scores for all families. However, gene trees are not data but can only be estimated based on a sequence alignment, for instance through maximization of Felsenstein (1981) likelihood. The species tree can then be obtained according to equation 4, and requires computing highest family likelihoods between gene and species trees, where both reconciliation and Felsenstein (1981) likelihoods are taken into account. To search for the highest family likelihood between a gene tree and a species tree, only the gene tree is modified. This modification can be obtained through commonly used tree search heuristics; for the sake of rapidity, we used a simple Nearest Neighbor Interchange (NNI) (Guindon and Gascuel, 2003) strategy, as follows:

1. For each branch of the current gene tree topology, a most likely reconciliation score is computed for the two possible NNIs.

2. If a reconciliation score is better than the best current reconciliation score, the reconciliation score is computed.

3. If a reconciliation score is better than the current reconciliation score, the NNI is accepted and the algorithm resumes with the new gene tree topology.

Step 1 in this algorithm implicitly assumes that the starting gene tree is the most likely according to Felsenstein (1981) likelihood, as it only computes the reconciliation score, which amounts to considering that only the reconciliation score can increase. In practice, starting trees can be obtained by PhyML (Guindon and Gascuel, 2003) for instance.

Finding the best species tree given several gene family trees: optimizing the full likelihood

The likelihood defined in eq. 4 can be used to compute the likelihood of a species tree and gene trees given sequence alignments. As the preceding sections have shown how one could compute the likelihood of a gene family, what remains to be explained to maximize this likelihood is how to explore the species tree topology as well as explore the space of parameters ruling the branch-specific probabilities of gene loss and gene duplication.

Finding the most likely species tree topology

To find the most likely species tree topology, classical algorithms can be used, with the simplification that the species tree does not have branch lengths, and the added difficulty that it needs to be rooted. The chosen algorithm is as follows:

1. For each subtree (node) in the species tree:

- prune the subtree
- regraft the subtree in all possible positions in the species tree, and compute the associated likelihood
- if such a Subtree Pruning And Regrafting (SPR) increases the likelihood, keep it

2. On each branch of the current species tree topology, root the species tree; as soon as a rooting improves the likelihood of the species tree, adopt it.
3. For each branch of the current species tree topology, perform all NNIs; as soon as a NNI improves the likelihood of the species tree, adopt it.
4. Iterate points 1 to 3 until no improvement is observed for a large number of steps.

Setting the branch-wise duplication and loss probabilities

In parallel to the species tree topology, values for the branch-wise rates of gene duplication and loss need to be found. This research can use the fact that branch-wise duplication and loss rates are independent of each other. Finding the most likely d_i and l_i for a branch i therefore only requires considering the counts of events on this branch i. To this end, one could use numerical optimization techniques, or use an analytical solution. For the sake of rapidity, we chose an approximate analytical solution. Normally, one should consider all counts of times where there were k lineages at the end of branch i, with $k \in [0; \infty[$; instead, we only use the numbers of times 0, 1 and 2 lineages have been found at the end of branch i. Using these counts only, approximate maximum likelihood values of d_i and l_i can be computed as follows:

$$\begin{aligned} d_i &= -\frac{ln(\frac{y+2z}{x+y+z})k(x+y+z)}{zx-yz-z^2+yx} \\ l_i &= -\frac{ln(\frac{y+2z}{x+y+z})x(y+2z)}{zx-yz-z^2+yx} \end{aligned}$$

where x, y, z are the numbers of times $0, 1, 2$ lineages are found at the end of branch i, respectively.

The algorithm as a whole

To search for the most likely species tree and simultaneously for the most likely gene family reconciliations under our model, we rely on a server-client architecture, schematized in Fig. 2.

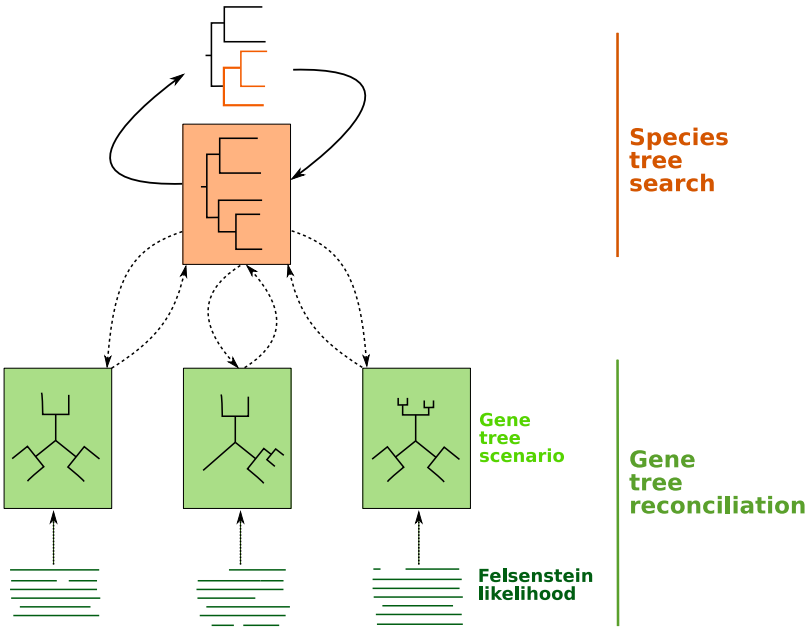


Figure 2: Server-Client architecture of the program. The server is in charge of the species tree search, as well as of the duplication and loss rates. It communicates with clients, each one in charge of one or more gene families, for which they build reconsiliations and compute family likelihoods, using sequence alignments.

The complete algorithm permitting to estimate the species tree simultaneously with gene trees is summarized in the following pseudo-code.

Algorithm 1 Optimizing the species tree as well as gene trees

likelihood_threshold=1e-6
$|T|$=2
if (server) {
    get initial Species tree (or build a random one) and store it into *currentS* and into *oldS*
    get the set of gene families to analyse
    create n clients
    send each client the set of gene families they are in charge of
    send each client *currentS* and rates of duplication and loss
    currentlk = -INFINITY
    while (iterations_without_improvement < limit) {
        receive family likelihoods and counts of gene duplications and losses from the clients
        compute total likelihood (newLk) for *currentS*
        if (newLk > currentLk)
            then {*oldS* = *currentS* ; iterations_without_improvement = 0 ; }
            else {*currentS* = *oldS* ; iterations_without_improvement ++ ; }
        change the topology of *currentS* and update rates of duplication and loss
        send each client *currentS* and rates of duplication and loss
    }
}
else if client {
    receive set of gene families
    receive *currentS* and rates of duplication and loss
    read alignments and PhyML pre-computed trees for each gene family
    while (iterations_without_improvement < limit) {
        compute family likelihoods
        send to the server family likelihoods and numbers of gene duplications and losses per species tree branch
        receive *currentS* and rates of duplication and loss
    }
}

Communications between the server and clients regarding the iterations_without_improvement variable are not shown.

Details of implementation

As the algorithm starts from a random species tree, reconciling gene family trees with it can take a very long time. To save computation time, during the first iterations, only reconciliation likelihoods are computed, *i.e.* gene family

trees are not modified. During these steps, branch-wise duplication and loss probabilities are the same for all the branches of the species tree, in order not to be trapped in a local maximum. Once the species tree has been optimized, a second optimization phase is started, where gene trees are modified, and values of duplication and loss rates are truly branch-wise. The program has been implemented with the help of the Bio++ (Dutheil *et al.*, 2006) and Boost libraries (Boost, 2008), and can run on clusters of computers using the Message Passing Interface (MPI).

Conclusion

We have built a program that can run on several computers in parallel to reconstruct a species tree simultaneously with gene trees in a maximum likelihood framework. Several modifications or improvements could easily be implemented.
First, all gene trees are considered to have branch lengths independent of the branch lengths of other genes. This is clearly unrealistic. Instead, one could use an approach similar to Rasmussen and Kellis (2007), by associating branch lengths to the species tree, and having one scaling parameter per gene family. This would also offer the highly desirable possibility to produce a dated species tree.
Second, better models of sequence evolution could be used. Noticeably, non-homogeneous models, that tolerate datasets showing heterogeneities in composition, could be used. With such models, different models of evolution could be associated to different branches of the tree, thus accounting for genome-wide biases. This addition would add another signal for rooting the species tree, a notoriously difficult problem, as non-homogeneous models are non-reversible.
Third, other events than gene duplications could be modelled, such as gene transfer and trans-specific polymorphisms (Wiuf *et al.*, 2004). Such events would only affect the reconciliation likelihood and would thus be easy to integrate.
Fourth, dependencies between gene trees could be added. Genes may share a history because they have remained close to each other on a chromosome throughout their history, in which case Hidden Markov Models as in Hobolth *et al.* (2007) may be useful, or because they interact as part of their function, in which case the more general framework of gene-to-tree maps (Ané *et al.*, 2007) may be relevant. Searching for correlations between genes may require larger changes in the programs however, as our algorithmic structure relies on the fact that genes can be considered as independent of each other.
Fifth, our program could easily be modified to compute Bayesian posterior probabilities instead of likelihoods, and to sample from posterior distributions through Markov Chain Monte Carlo distributions instead of providing maximum likelihood estimates.
We believe that approaches such as this are going to be very useful in the future. Distinguishing the species tree from gene trees may be costly but is a necessary step towards a proper estimation of the tree of life.

References

Ané, Cécile, Larget, Bret, Baum, David A, Smith, Stacey D, and Rokas, Antonis. 2007. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, **24**(2), 412–426.

Arvestad, Lars, Berglund, Ann-Charlotte, Lagergren, Jens, and Sennblad, Bengt. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, **19 Suppl 1**, i7–15.

Bansal, Mukul S, and Eulenstein, Oliver. 2008. The multiple gene duplication problem revisited. *Bioinformatics*, **24**(13), i132–i138.

Blanquart, Samuel, and Lartillot, Nicolas. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol Biol Evol*, **23**(11), 2058–2071.

Blanquart, Samuel, and Lartillot, Nicolas. 2008. A Site- and Time-Heterogeneous Model of Amino-Acid Replacement. *Mol Biol Evol*, Jan.

Boost. 2008. *Boost C++ libraries*. http://www.boost.org/.

Boussau, Bastien, and Gouy, Manolo. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol*, **55**(5), 756–768.

Chen, K., Durand, D., and Farach-Colton, M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol*, **7**(3-4), 429–447.

Ciccarelli, Francesca D, Doerks, Tobias, von Mering, Christian, Creevey, Christopher J, Snel, Berend, and Bork, Peer. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**(5765), 1283–1287.

Degnan, James H, and Rosenberg, Noah A. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet*, **2**(5), e68.

Delsuc, Frédéric, Brinkmann, Henner, and Philippe, Hervé. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, **6**(5), 361–375.

Delsuc, Frédéric, Brinkmann, Henner, Chourrout, Daniel, and Philippe, Hervé. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**(7079), 965–968.

Doolittle, R. F., and Blombaeck, B. 1964. AMINO-ACID SEQUENCE INVESTIGATIONS OF FIBRINOPEPTIDES FROM VARIOUS MAMMALS: EVOLUTIONARY IMPLICATIONS. *Nature*, **202**(Apr), 147–152.

Dubb, Lindsey. 2005. *A Likelihood Model of Gene Family Evolution*. Ph.D. thesis, University of Washington.

Dunn, Casey W, Hejnol, Andreas, Matus, David Q, Pang, Kevin, Browne, William E, Smith, Stephen A, Seaver, Elaine, Rouse, Greg W, Obst, Matthias, Edgecombe, Gregory D, Sørensen, Martin V, Haddock, Steven H D, Schmidt-Rhaesa, Andreas, Okusu, Akiko, Kristensen, Reinhardt Møbjerg, Wheeler,

Ward C, Martindale, Mark Q, and Giribet, Gonzalo. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**(7188), 745–749.

Dutheil, Julien, Gaillard, Sylvain, Bazin, Eric, Glémin, Sylvain, Ranwez, Vincent, Galtier, Nicolas, and Belkhir, Khalid. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, **7**, 188.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, **17**(6), 368–376.

Felsenstein, J., and Churchill, G. A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol*, **13**(1), 93–104.

Felsenstein, Joe. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*, **27**, 401–410.

Foster, Peter G. 2004. Modeling compositional heterogeneity. *Syst Biol*, **53**(3), 485–495.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol*, **18**(5), 866–873.

Galtier, N., and Gouy, M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol*, **15**(7), 871–879.

Goodman, M.J., Czelusniak, G., Moore, G., Romero-Herrera, A., and Matsuda, G. 1979. Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, **28**, 132–168.

Gowri-Shankar, V., and Rattray, M. 2006. On the correlation between composition and site-specific evolutionary rate: implications for phylogenetic inference. *Mol. Biol. Evol.*, **23**(2), 352–364.

Gowri-Shankar, V., and Rattray, M. 2007. A Reversible Jump Method for Bayesian Phylogenetic Inference with a Nonhomogeneous Substitution Model. *Mol. Biol. Evol.*, **24**(6), 1286–1299.

Guigo, R., Muchnik, I., and Smith, T. F. 1996. Reconstruction of ancient molecular phylogeny. *Mol Phylogenet Evol*, **6**(2), 189–213.

Guindon, Stéphane, and Gascuel, Olivier. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52**(5), 696–704.

Hobolth, Asger, Christensen, Ole F, Mailund, Thomas, and Schierup, Mikkel H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet*, **3**(2), e7.

Huelsenbeck, John P, Bollback, Jonathan P, and Levine, Amy M. 2002. Inferring the root of a phylogenetic tree. *Syst Biol*, **51**(1), 32–43.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. *Pages 21–132 of:* Munro, M.N. (ed), *Mammalian Protein Metabolism*, vol. 3. Academic Press New York.

Kumar, Sudhir. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet*, **6**(8), 654–662.

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

Liolios, Konstantinos, Mavromatis, Konstantinos, Tavernarakis, Nektarios, and Kyrpides, Nikos C. 2008. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*, **36**(Database issue), D475–D479.

Mirkin, B., Muchnik, I., and Smith, T. F. 1995. A biologically consistent model for comparing molecular phylogenies. *J Comput Biol*, **2**(4), 493–507.

Page, R. D., and Charleston, M. A. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*, **7**(2), 231–240.

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

Rasmussen, Matthew D, and Kellis, Manolis. 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res*, **17**(12), 1932–1942.

Thorne, J. L., Kishino, H., and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol*, **33**(2), 114–124.

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

Wehe, André, Bansal, Mukul S, Burleigh, J. Gordon, and Eulenstein, Oliver. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, **24**(13), 1540–1541.

Weisburg, W. G., Giovannoni, S. J., and Woese, C. R. 1989. The Deinococcus-Thermus phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst Appl Microbiol*, **11**, 128–134.

Wiuf, Carsten, Zhao, Keyan, Innan, Hideki, and Nordborg, Magnus. 2004. The probability and chromosomal extent of trans-specific polymorphism. *Genetics*, **168**(4), 2363–2372.

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 Roberts, D. 1995. On the Use of Nucleic Acid Sequences to Infer Branchings in the Tree of Life. *Mol. Biol. Evol.*, **12**(3), 451–458.

Yap, Von Bing, and Speed, Terry. 2005. Rooting a phylogenetic tree with non-reversible substitution models. *BMC Evol Biol*, **5**(1), 2.

Zmasek, C. M., and Eddy, S. R. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**(9), 821–828.

Zuckerkandl, Emile, and Pauling, Linus. 1962. *Horizons in Biochemistry.* Academic Press, New York. Pages 189–225.

10

Problems and Perspectives for the Evolutionary Study of Genomes

My thesis has addressed issues related to phylogenetic reconstruction, as well as issues related to the reconstruction of the mechanistics of evolution. Both fields are expected to make much progress in the next years, and this last article attempts to foresee what these advances will be, by analysing recent literature.

This article has not been submitted yet.

Genomes as documents of evolutionary history

Bastien Boussau Vincent Daubin

October 22, 2008

Université de Lyon ; université Lyon 1 ; CNRS ; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France.

Contents

Abstract

As primary semantides (Zuckerkandl et Pauling, 1965), genomes conceal a vast intricate record of their carriers descent and evolution. To unravel this information, Phylogenetics must assimilate all aspects of genome evolution. In return, much biology can be learned. The emerging field of Phylogenomics incarnates this power of genomics and evolution to mutually highlight each other. New approaches integrating population genetics, genome evolution, geography, geology and/or ecology into phylogenetic models are now emerging and arguably anticipate the future of this discipline. In this article, we review recent advances, and discuss possible developments towards a comprehensive reconstruction of the history of life.

1 Introduction

Like many brilliant ideas, the tree of life is a simple concept. When Charles Darwin, building on a theme initiated by his precursor evolutionists, elaborated upon the metaphor of a tree *"which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications"* (Darwin, 1859), he was probably picturing a rather plain, bifurcating, occasionally multifurcating pattern of relationships among species. No doubt that, being familiar with the tendency of living systems to produce exceptions to most rules, he would have been able to further anticipate the existence of complexifying paths such as chimerism and endosymbiosis. However, the difficulty of the task of reconstructing a fully resolved and dated history of life could hardly be foreseen only a few decades ago. The breadth of life's diversity, the interwoven routes of evolution, and the complexity of phylogenetically operational characters still hamper attempts of evolutionists to produce a comprehensive picture of biological evolution. Sometimes to the point of resignation (Doolittle, 1999; Rokas et Carroll, 2006). And in the current deluge of genome sequences, it has become more and more difficult to retreat arguing that "more data are needed".

Long before Zuckerkandl and Pauling proposed that genes could be used as documents of evolutionary history (Zuckerkandl et Pauling, 1965), Sturtevant and Dobzhansky (Sturtevant et Dobzhansky, 1936) had already built what is to our knowledge the first genome phylogeny: a tree of *Drosophila pseudoobscura* strains, based on the parsimonious reconstruction of large chromosomal inversion scenarios. Since then, many different approaches have been devised to exploit the information contained in genomes and build species trees (see Snel *et al.* (2005); Delsuc *et al.* (2005) for recent reviews). Today that we have hundreds of complete genome sequences, the results of these phylogenomic approaches are mixed: several "highly resolved" Trees of Life have been published that conflict with each other (Brown *et al.*, 2001; Battistuzzi *et al.*, 2004; Ciccarelli *et al.*, 2006), and their interpretation has been controversial (Bapteste *et al.*, 2005; Dagan et Martin, 2006). However, this apparent dead end could be escaped by picking a few methodological locks. The recent litterature is flourishing with first attempts at coupling inference of species and gene phylogenies with models of population genetics or genomic evolution. These approaches pioneer the truly integrative science that phylogenomics ought to be: a science that combines various levels of complexity and different fields of evolutionary biology. In this paper, we propose to review specifically these recent breakthroughs. We will also discuss and try to anticipate the further developments that are needed to reconstruct a comprehensive history of life.

2 The "orthologous gene family" myth

Why have phylogenomics methods been unsuccessful at producing an undisputed tree of life? Although abundant, molecular characters appear of considerable complexity for phylogenetics. Coined by Walter Fitch (Fitch, 1970), the term "ortholog" designates genes that are related through speciation events, as opposed to "paralogs", which are the result of duplications. Therefore, with the intent of reconstructing a phylogeny of species, the interpretation of trees based on orthologous genes should be straighforward. But the combined effect of hidden paralogies, lateral gene transfer (LGT), trans-specific polymorphism (TSP) and phylogenetic artifacts (Box 1) makes the process of reconstructing the history of species difficult.

A common opinion in the field of phylogenomics is that data are so abundant that one can generously cut into them to get an ideal dataset. Indeed, one of the reasons why current approaches have been yet unsuccessful at producing a comprehensive view of the evolution of life, is precisely because they are usually

not comprehensive. When reconstructing the tree of life through the combination of gene families, one usually choses those genes having representatives in most species under study, and showing no obvious evidence for complexifying events in their histories such as duplication or LGT. To infer deep phylogenetic relationships, only a handful of genes are therefore used, with the hope that the phylogenetic signal for the species tree will prevail. However, combining data in the presence of LGT, paralogy or TSP can be positively misleading (Brown *et al.*, 2001; Degnan et Rosenberg, 2006). Trying to first remove conflict among combined datasets further decreases the number of genes under study, making the resulting tree appear as an anecdotal picture of the history of life.

At the other extremum, exhaustive gene repertoires have been used for phylogenetic inference, but at the price of ignoring the phylogenetic information carried by sequences to focus on the presence and absence of genes in genomes. However, not considering gene histories seems unsound when we know so little about the probabilities of gene transfers and losses, and the models of evolution applied for such reconstructions are usually overly simple, with all genes having the same probability of being acquired or lost on the entire tree. Although other rare genomic changes (RGC) (Rokas et Holland, 2000) have been used for phylogenetic inference of deep phylogenies, most of them are expected to be just as sensitive to hidden paralogy and lateral gene transfer.

The impact of most processes described in Box 1 is only expected to increase with more data. So what can we do, when we come to admit that, in spite of the deluge of genomic sequences, there is no and will never be a perfect dataset, devoid of lateral gene transfer, incomplete lineage sorting, hidden or apparent paralogies, convergent gene losses, systematically biased or accelerated evolutionary rates *etc.*..? The answer is probably: exploit the evolutionary significance of these events. A proper modelling can not only improve phylogenetic reconstruction, but also bring further insight into the evolution of life: lateral gene transfer may for instance provide strong support for the monophyly of some groups of species and unvaluable information about the relative timing of clade diversification and ecological affinities; duplications and losses being the hallmarks of genome dynamics, can be used to better understand the relationships between genomes structure and species diversification or ecology; incomplete lineage sorting offers previously unforeseen opportunities to estimate ancestral population sizes and divergence times. Interestingly, these three sorts of phenomena can be modelled similarly.

3 Phylo-gene-ethics

The reconstruction of gene histories traditionally relies solely on a gene alignment, and a model of nucleotide or amino-acid substitution. There are however additionnal constraints that can be enforced to a gene tree during the process of reconstruction: the most obvious biological knowledge that can be used to improve gene tree reconstruction is the fact that every gene evolves within the banks of a species phylogeny. According to this view, a gene tree is a deformation of the species tree through the prism of the evolutionary events described above (Fig. 1). If we are able to correctly model the different processes that make genes and species tree differ and if several genes that have evolved under this same constraint can be treated together, then gene and species trees can be searched simultaneously, which should result in better trees as well as increased knowledge of the processes (Maddison, 1997; Suchard *et al.*, 2003) (Fig. 2).

Estimating genes and species history can thus be achieved through a hierarchical structure, on top of which a species tree is inferred from gene trees through models of gene family evolution, themselves inferred from sequence alignments through models of sequence evolution (Fig. 2). The relationship between the species tree and gene trees is two-ways: the species tree induces a probability distribution over gene trees (some gene trees are more likely than others given a particular species tree), and in return, the inferred distribution of gene trees informs about the species tree along which they were generated.

Considering that genes evolve in the context of a species tree should result in better gene trees. Indeed, there are several reasons why a reconstructed gene tree may differ from the true gene tree: gene sequences may have undergone too many substitutions (possibly leading to long branch attraction), sequence compositions may widely differ from one gene to the other (compositional heterogeneity), or gene sequences may be so constrained or so small that there is simply not enough signal in these to reconstruct their history. In such cases, classical phylogenetic methods will output gene trees that may be very different from the true trees. Injecting a species tree into the process of inference can counterbalance reconstruction artifacts, and should thus result in better gene trees.

Figure 1: Gene trees are deformed reflections of the species tree that constrained their evolution. Credit Françoise Boussau

Three natural processes may produce gene trees different from species trees: trans-specific polymorphisms, duplications, and Lateral Gene Transfers (Box 1). Consequently three different models of gene family evolution can be devised to model gene family evolution. In the next paragraphs, we will present models and algorithms that have been developed to account for the processes of gene family evolution.

4 Population and species histories

It might not be obvious *a priori* why processes acting at the population level influence species phylogenies. Recent work (Degnan et Rosenberg, 2006) however showed that, in some conditions of population sizes, population structures and divergence times, most gene trees differ from the species tree, and concatenating these genes will converge to an incorrect estimate of the species tree. Although the impact of population genetics processes on gene tree topologies may not always be so severe, coalescent theory (Kingman, 1982) predicts that under neutral evolution a proportion of gene trees will differ from the species tree through TRANS-SPECIFIC POLYMORPHISMS (TSP) (Fig. 1). More precisely, if the number of generations separating two speciations is not very large compared to the population size observed between these two speciations, it becomes likely that the coalescence of genes present in two species is more ancient than the previous speciations, which results in a gene tree different from the species tree. The amount and types of gene tree / species tree incongruences thus inform about divergence times and ancestral population sizes. Models of population genetics improve the reconstruction of a species tree, and may even be the only way to get a correct species tree in some cases (Degnan et Rosenberg, 2006; Kubatko et Degnan, 2007; Rosenberg et Tao, 2008).

Models of TSP based on the coalescent framework (Kingman, 1982) have been proposed several times (Beerli et Felsenstein, 1999; Rannala et Yang, 2003; Maddison et Knowles, 2006; Liu et Pearl, 2007; Carstens et Knowles, 2007). The first models assumed a known species phylogeny and estimated divergence times along

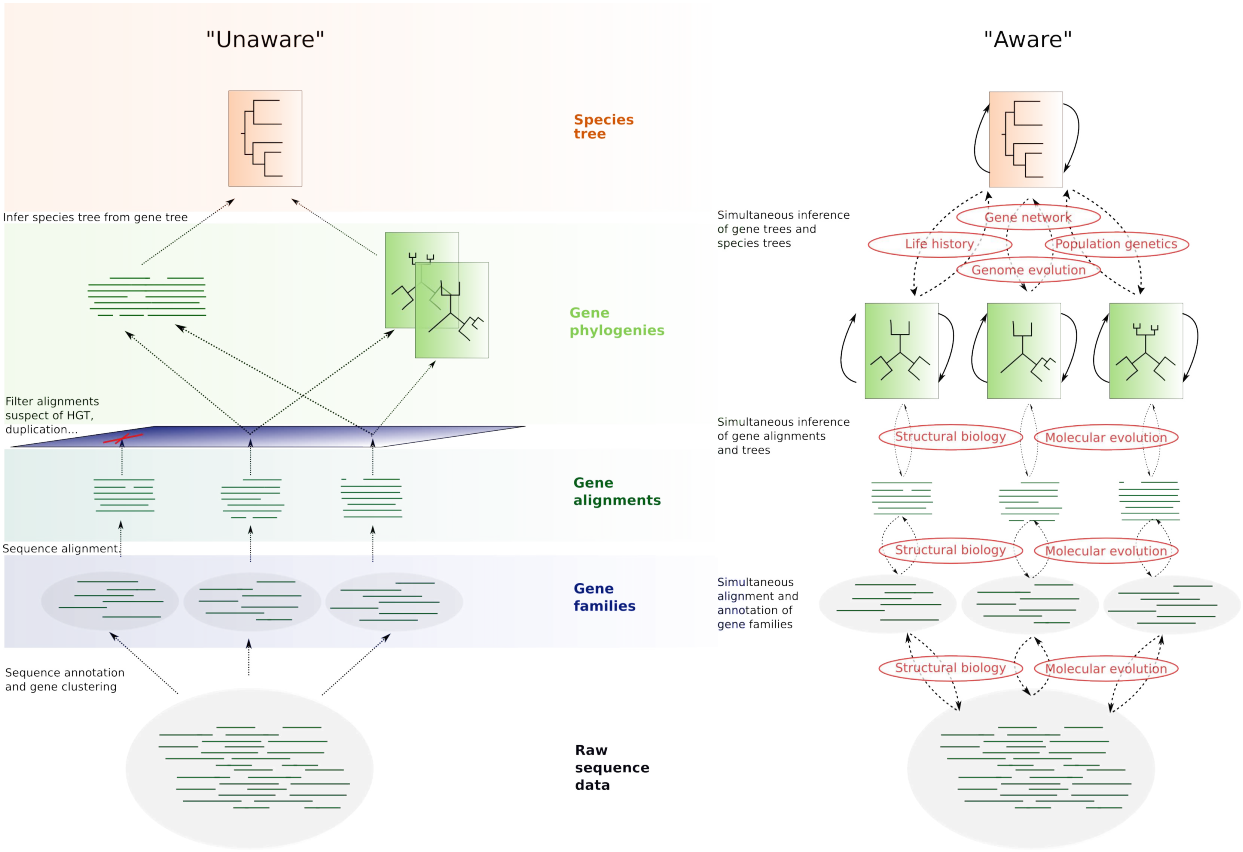


Figure 2: Phylogenetic awareness: the two paths from sequences to species tree. In the "unaware" path (the traditional way of inferring species phylogenies) each stage of the phylogenetic inference is essentially independent from the steps up- and downstream. In addition, sequence alignments have to pass different filters in order to make gene trees readily understandable as species trees (abscence of duplicates, LGT...). In contrast, the "aware" path models the dependency between each step and degree of complexity using knowledge from different fields of biology (red ellipses, the list is not exhaustive): alignments can be statistically estimated simultaneously with gene trees, using models of sequence evolution that incorporate insertion/deletion events; and models of gene family evolution incorporating LGT, duplication and/or incomplete lineage sorting specify the dependency between gene trees and species tree. Two ways arrows represent these dependencies, and continuous arrows represent gene tree and species tree searches. The dependency between gene family annotation, alignment and phylogenetics has not been yet explored, but could theoretically be modelled (see text for discussion). The schematic representation of the synchronous search for species trees, gene trees, gene alignments etc... suggests an obvious achitecture for parallelizing this search.

with ancestral population sizes (Beerli et Felsenstein, 1999; Rannala et Yang, 2003; Maddison et Knowles, 2006); more recent ones can also estimate the species phylogeny (Liu et Pearl, 2007; Carstens et Knowles, 2007). Up to now, no approach has inferred both gene trees and species trees simultaneously, although this would be the natural but costly approach to take. For instance, Carstens et Knowles (2007) first estimated gene trees in the maximum likelihood framework from sequence alignments; they then used these trees to compute the likelihoods of candidate species tree topologies according to the coalescent model. All evaluated species trees were then compared using likelihood ratio tests. Simulations showed that this approach considerably outperformed concatenation: species trees were correctly reconstructed using the coalescent-based approach even for very recent and close speciations, whereas concatenation trees were most often different from the true species tree in the same conditions. Instead of maximum likelihood, Liu et Pearl (2007) used

a Bayesian framework through MARKOV CHAIN MONTE CARLO (MCMC) sampling. They went a step further towards the joint recontruction of gene trees and species trees, by recognizing that if species tree and ancestral population sizes can be estimated based upon gene trees, they also influence gene trees in return. Their approach is composed of three steps: first, gene trees are reconstructed using MrBayes (Huelsenbeck et Ronquist, 2001), approximately accounting for an unknown species tree; second, a species tree is inferred based on the distribution of gene trees obtained previously; third, the approximation made during the first step is corrected. In the end, for each sequence alignment, a distribution of gene trees is obtained, as well as distributions of species trees and ancestral population sizes. As for the formerly described approach, this method was found to have a much better fit to the data than gene concatenation; moreover this Bayesian approach is superior to the maximum likelihood one in several respects, as it can analyse a larger number of species, allows different ancestral population sizes for different nodes of the species tree, and can estimate all parameters of the model.

Several confounding factors from population genetics however may need to be modelled to get correct estimates of species trees, gene trees, divergence times and ancestral population sizes. Notably, selection will affect the gene trees in a way that can be misinterpreted in terms of population size. Balancing selection, for instance, will favor TSPs in a manner virtually independent from population size and mimic the evolution of a neutral gene in a large population; strong purifying selection on the other hand will mirror trees obtained on neutral loci in small populations. Similarly, migration will also introduce discording gene trees, and therefore calls for more complex models (Innan et Watanabe, 2006).

5 Duplications and losses

The combined action of gene duplication and loss considerably complexifies gene trees (Fig. 1). Even a gene family with only one representative per species may harbour events of duplication and loss, yielding a gene tree different from the species tree. Models of gene family evolution taking into account duplications and losses have been proposed by Arvestad *et al.* (2003) and Dubb (2005). In both cases, the evolution of a gene family is modelled by a birth-death process running along a species tree: "birth" corresponds to gene duplication, and "death" to gene loss. If in their original model, both gene trees and species trees were fixed, Arvestad and co-workers (Arvestad *et al.*, 2004) later combined their model of gene family evolution with a model of sequence evolution, so that given a species tree, likelihoods of gene trees can be computed. The model has been implemented in a program that can estimate gene trees and gene orthology/paralogy probabilities given a species tree, through Bayesian MCMC integration. The estimation of a species tree however was only tackled theoretically, for computational reasons: the program therefore needs a user-input species phylogeny, but can compute gene trees that are biologically more meaningful than if they had been inferred based on a sequence alignment alone. In addition, this model provides posterior probabilities of orthology and paralogy for each pair of genes, which could be further used as an aid for functional prediction. However, to date, these models use a single duplication probability and a single loss probability for all branches of a tree, even though it is known that different lineages undergo different rates of duplication and loss. Future models should cope with this inhomogeneity of the evolutionary process to properly depict gene family evolution. Moreover, relying on a known species tree for building accurate gene trees is certainly optimistic in several cases: just as Arvestad *et al.* (2004) program integrates over scenarios of duplications and losses with respect to a given species tree, a better statistical estimation of gene tree and of orthology/paralogy probabilities should be obtained by integrating over the distribution of species trees. Such a model would provide a species tree built using more than the genes that happen to be single-copy in most genomes and may help resolve some difficult phylogenies (Dunn *et al.*, 2008), and at the same time clarify the dynamics of genome expansion/shrinking over the entire tree.

6 The reticulated tree

When modelling lineage sorting and duplication, it seems reasonable to consider that there exists an underlying species tree, *i.e.* a tree depicting the history of vertical inheritance in the genome. In the case of Lateral Gene Transfer however, the issue of whether the concept of a species tree applies has been discussed

at length (Doolittle, 1999; Ochman *et al.*, 2000; Kurland *et al.*, 2003; Daubin *et al.*, 2003; Lerat *et al.*, 2003). Today, it seems clear that the phylogenetic information of gene trees is structured in a way that suggests at least some vertical signal in the history of life, but the debate still doesn't appear to be settled. The current confusion on this topic might be due only to inappropriate methods. Indeed, most gene families have had time, during their history, to be transferred among distant organisms. And if one could find a family devoid of any evidence of transfer today, further sampling of genetic diversity would certainly bring some. However, does it mean that there is no vertical signal to be found in gene trees? An appropriate modelling of LGT would probably help to clarify the issue.

Suchard (2005) reached a first stage in the elaboration of such a model. In his model, a species tree produces a distribution of gene trees through topological rearrangements mimicking LGT, and gene trees are evaluated with respect to gene alignments. The whole model was implemented in the Bayesian framework with MCMC sampling. The analysis produced further evidence for the "complexity hypothesis", that genes which are part of large protein complexes are less subject to LGT; however, the dimension of the topological space that can be reached through LGT simulation from a single species tree limited the applicability of the method to six species. Other approaches using fast algorithms reconciliating gene and species trees under a LGT model may help tackle this problem (Addario-Berry *et al.*, 2003; Hallett *et al.*, 2004). The result of a program implementing such a model would be a species tree decorated with LGT frequencies, and would thus permit to formally test the "tree of life" hypothesis: are the frequencies of transfers found on each branch such that no tree should be preferred over all the other possibilities? Are there a few tree topologies that are found significantly more often than others? Can we relate these topologies to ancient events of endosymbiosis or ecological shifts? Are species sharing similar ecological niches exchanging large numbers of genes? Are different types of genes differentially transferred? These questions are the object of an impressive amount of work (Beiko *et al.*, 2005; Ge *et al.*, 2005; Zhaxybayeva *et al.*, 2006; Choi et Kim, 2007; Dagan et Martin, 2007) that would strongly benefit from a proper statistical framework, that does not take gene trees as data, but as statistical estimates from gene sequences themselves. In addition, gene transfer events constitute informative characters for phylogenetic reconstruction (Huang et Gogarten, 2006), and provide relative dates for nodes of a species tree: if a descendant from node A gives a gene to an ancestor of node B, this means that node B is more recent than node A. Considering the immense difficulty of dating nodes in the prokaryotic tree of life where fossils are scant and at best difficult to compare to extant species, such a relative dating would certainly be highly valuable.

7 Computational challenges

The above-described models did not infer at the same time species tree, gene trees and parameters of the model of gene family evolution, or did so for a very modest number of species (Suchard, 2005). In fact, the hierarchical structure including models of gene family evolution (Fig. 2) represents an interesting challenge for algorithmics. Searching for a gene tree based on an alignment is already an intimidating task, as the number of topologies increases more than factorially with the number of leaves; searching for both a species tree and several gene trees at the same time is even more difficult: for each species tree, one needs to estimate/sample corresponding gene trees. This computational space explains why no program has yet been devised that could efficiently infer both gene and species trees simultaneously. There is however an obvious way to parallelise computations through an architecture based on a server and several clients: a server node would be searching for a species tree, while client nodes would, for each species tree, search corresponding gene trees. Such a parallelisation would be necessary to compute trees based not on a few genes for a few species, but on whole genomes, which may require hundreds or thousands of computers running for several days.

The most complete model of gene family evolution would account for gene transfers, duplications, losses and discrepancies in lineage sorting, but may be difficult to devise while avoiding overparametrization (Box 2). However, progress in algorithmics may render such models computationally tractable (Hallett *et al.*, 2004; Than *et al.*, 2007), and allow inference of the relative contributions of these events to the large amounts of incongruences observed in Prokaryotes for instance. To be biologically realistic, such models should associate different probabilities of gene duplication, loss and transfer to different branches of the species tree. If modelling all three types of events for a given branch turns out to be too difficult, an economic alternative

would allow different kinds of events for different parts of the tree, or more elegantly automatically affect models to regions of the tree.

Even a model incorporating trans-specific polymorphisms, LGT, and duplication/loss may be over-simplistic if it does not account for dependencies between genes. For instance, two neighboring genes, from a bacterial operon or from a Eukaryotic chromosome, are more likely to share a similar history than genes from different regions of the genome. Coevolution may also affect genes that interact as part of their function (Barker et Pagel, 2005; Sémon et Wolfe, 2007a; Aury *et al.*, 2006). Accounting for effects of spatial proximity on gene histories can be achieved through Hidden Markov Models (HMM), as recently used by Hobolth *et al.* (2007) to infer recombination hotspots simultaneously with ancestral population sizes and divergence times. Another more general approach to model coevolution may use gene-to-tree maps (Ané *et al.*, 2007), with adequate models of dependency. Gene-to-tree maps are objects which associate genes with distributions of trees: two genes that have co-evolved for a part of their history will show partially similar tree topologies. How similar their topologies will be depends on the type of coevolution, and this can be injected into a statistical model through prior probabilities: for instance, two interacting genes are *a priori* more likely to share tree topologies than genes that are part of two totally different pathways.

Overall, accounting for the evolutionary processes acting at the gene family level is mandatory if one wants to avoid bias in estimating a species tree and gene trees; additionally, proper models of gene family evolution could provide more information than the mere pattern of species diversification, such as divergence times and ancestral species effective population sizes. Accurately reconstructing gene trees would even offer new possibilities to study the evolution of genome contents: up to now, genome content reconstructions only relied on counts of genes (Snel *et al.*, 2002; Boussau *et al.*, 2004; Hao et Golding, 2008). Using gene trees instead would probably greatly improve inferences.

8 The alignment layer

No matter how sophisticated models of gene family evolution may get, if the alignments associated with gene families are incorrect, so will probably be the resulting gene and species trees. The classical textbook representation of phylogenetic reconstruction is a three step process: first homologous genes are selected, then aligned (the phylogeneticist may intervene here and esthetically polish the alignment), and finally a tree is built from the alignment (Fig. 2). This view however amounts to considering a sequence alignment as data whereas it is an estimate: most alignment programs use heuristics to place gap characters into sequences (lines) so as to organize putatively homologous sites into columns. The optimality of gap placement is assessed with respect to a score, which penalizes gap insertions, gap extensions, and substitutions. In the end, the alignment is the best estimate of the true alignment, according to arbitrary penalties which may be unrealistic for the data under study, and according to a particular heuristics (Thompson *et al.*, 1994; Notredame *et al.*, 2000; Edgar, 2004; Löytynoja et Goldman, 2008), which even if the penalties were perfectly tuned to the data, may not find the optimum alignment; still, if the optimum alignment is found, there is no guarantee that it is the true alignment. These sobering considerations have long been known, and the limitations inherent to relying on a single alignment to infer a phylogenetic tree are now well accepted (Lake, 1991; Landan et Graur, 2007; Wong *et al.*, 2008). However, only recently have people tried to solve the problem with a statistically sound approach, first with a probabilistic model of insertion and deletion events combined with classical substitution matrices, and second by jointly estimating sequence alignments and phylogenetic trees.

The first probabilistic model for the maximum likelihood alignment of two sequences was devised by Bishop et Thompson (1986); a more realistic model was later proposed by Thorne *et al.* (1991), although it considered only point insertions and deletions between two sequences. A year later, it was enriched to account for multiple-site insertions and deletions (Thorne *et al.*, 1992). Such models of statistical alignment can be seen as relying on Hidden Markov Models (or on the closely related transducers (Bradley et Holmes, 2007)), in which states are "match", "insertion", and "deletion". More recently, as Bayesian methods have become increasingly popular, several algorithms implementing Bayesian MCMC joint samplings of multiple gene alignments and gene trees have been proposed (Mitchison, 1999; Hein, 2001; Holmes et Bruno, 2001; Metzler, 2003; Lunter *et al.*, 2005; Redelings et Suchard, 2005). Associating pairwise alignments based on HMMs to each branch of a phylogenetic tree permits to easily compute the likelihood of a multiple alignment (Holmes et Bruno, 2001), but integrating over the distribution of probable alignments and trees is

very computationally intensive. However, it appears as the best approach for estimating phylogenetic trees while accounting for all the uncertainty in an alignment.

Moreover, Bayesian joint estimations of sequence alignments and phylogenetic trees offer the possibility to better characterize the sequence evolutionary process, as probabilities of insertions and deletions can be simultaneously estimated and compared with substitution probabilities. Contrary to most commonly used software packages, programs that simultaneously estimate alignments and phylogenies do not treat gaps as unknown characters, but can use insertion-deletion as phylogenetically informative events. In the case of SIV and HIV viruses, this has been shown to improve resolution of the phylogenetic tree (Redelings et Suchard, 2007); as the rate of insertion/deletion is believed to be lower than substitution rates, their incorporation into phylogenetic reconstruction may also help resolve ancient divergences. Moreover, as a deleted character cannot be reinserted (otherwise it is not considered homologous), insertion-deletion events can impose a direction on a phylogenetic tree, and therefore point to its root. Since rooting a phylogenetic tree is notoriously difficult (Huelsenbeck *et al.*, 2002; Yap et Speed, 2005), insertion-deletion events contain information worth exploiting, for rooting gene trees, but also for rooting species trees. In this respect, working on a four-level hierarchic model that would go from sequences to species trees through alignments and gene trees (Fig. 2) may not harm too much the computational efficiency of alignment sampling. Indeed, the analysis of a single alignment may be difficult in part because there is not much information contained in its sequences: there may therefore be a lot of uncertainty on the phylogenetic tree, so that in a Bayesian setting, a large number of gene trees need to be visited, and for each of these trees a large number of alignments need to be sampled. However, if several genes were aligned in parallel, each gene would benefit from the information of other genes through their common species tree: the distribution of probable gene trees would be narrower, and in consequence it might take less time to jointly sample alignments and trees.

In addition to phylogenetic reconstruction, other sequence-based inferences can benefit from averaging out the alignment. For instance, the detection of sites under positive selection has recently been shown to depend upon the method of alignment used (Wong *et al.*, 2008), and the reliablility of protein structure prediction is inversely correlated to alignment ambiguity (Miklós *et al.*, 2008): using a single alignment to predict protein structure is therefore likely to lead to unproper inferences in portions of proteins that are uneasy to align. Similar conclusions have been drawn for phylogenetic footprinting techniques. These methods benefit from the comparison of several genomes to detect putatively functional regions, *i.e.* portions of sequences that are more conserved than expected and therefore must be under purifying selection (Duret *et al.*, 1993). Their relying on a single alignment however affects the quality of inferences: when two different alignment algorithms were used to annotate transcription factor binding sites in 12 drosophila genomes, they agreed upon less than 60% (Stark *et al.*, 2007). Consequently, Satija *et al.* (2008) devised an algorithm to detect slowly evolving regions where the alignment was not fixed but integrated over, and showed that this integration could significantly improve binding site detection when the binding sites were not perfectly conserved.

9 Recombination and homology

As if the task of reconstructing gene and species histories was not complicated enough, considering gene families as the unsecable bricks of phylogenetic reconstruction is incorrect. The shuffling of genetic material, through the processes of recombination and gene fusion frequently produces genes with mixed phylogenetic signals or even heterologous parts. Homologous recombination, the replacement of part of a sequence by a relative will simply result in gene alignments with contadictory phylogenetic signal over their length. Gene fusion will have a deeper impact, and affect the earlier steps of inference of gene homology, as only parts of protein sequences can be considered homologous. In any case, the reconstruction of a gene family history based on its entire length may be at best partial, or completely wrong. Although the only truly irreducible homologous character is the nucleotide, these events may comprise relatively long streches of sequence, and the conflicting signals can be identified.

Many approaches have been devoted to identifying events of homologous recombination in multiple gene alignments, and recent models can simultaneously search for segment boundaries and histories in an alignment (Minin *et al.*, 2005; Kedzierska et Husmeier, 2006; Pond *et al.*, 2006). Although this step has not been undertaken yet, a species tree reconstruction model using multiple genes could be devised which introduces

these models of gene recombination into the LGT model described above.

Gene fusion and domain shuffling will probably be more complicated to model as they have an impact on the primary hypothesis of homology, *i.e.*, the attribution of a protein to a family. Protein homology is typically inferred from their overall similarity and several public databases propose automatically reconstructed gene families based on this criterion. Although the clustering method used to group proteins may vary, local similarities are usually dismissed, and protein sequences sharing homologous segments can be typically placed in different (heterologous) families. The high significance of this protein modularity, which can be viewed as a "level of homology" problem, can be gauged from the fact that there may be 19% of eukaryotic exons that have undergone recombination with a non-homologous portion of the genome (Long *et al.*, 2003). Interestingly, compared to entire proteins, domains should provide information on deeper phylogenetic relationships. An ideal way of dealing with this issue would be to couple the processes of homology assignment, sequence alignment and phylogenetic reconstruction into a model able to reconstruct and combine trees at different levels of homology.

10 Genome alignment

Reconstructing the evolution of genomes is not only reconstructing the history of their genes. Events such as inversions, tandem duplications, chromosome fission/fusion also affect genomes, and need therefore to be modelled to properly depict their evolution. However, the consideration of such events considerably complexifies the alignment problem, and consequently most published approaches have been based on parsimony. Recently, Larget *et al.* (2005) devised a Bayesian program to estimate distributions of ancestral genome arrangements as well as the genome phylogeny in 87 animal species, using all the 37 genetic markers available in the mitochondrial genome. The underlying model considered only inversions as possible rearrangements. Using the same program Darling *et al.* (2008) analysed the distribution of rearrangement scenarios and finely characterized the dynamics of 8 genomes of closely related bacterial strains, containing 78 conserved groups of genes.

Devising a model of genome evolution taking into account inversions, duplications, losses, chromosome fission/fusion would probably offer more insight into genome evolution, but constitutes a considerable challenge; further incorporating substitutions and insertions/deletions would build a model able to align whole genomes in a statistical framework, and should result in improved understanding of genome dynamics, as it is known that the genomic environment influences substitution pattern (Eyre-Walker et Hurst, 2001; Daubin et Perrière, 2003; Lobry, 1996; Necsulea et Lobry, 2007). This may help test theories of speciation by chromosomal rearrangements (Rieseberg, 2001), help detect and characterize whole genome duplications (Sémon et Wolfe, 2007a), and pave the way for the reconstruction and study of ancestral genomes (Sturtevant et Dobzhansky, 1936; Burt *et al.*, 1999; Muffato et Crollius, 2008). Moreover, such approaches would offer an integrative framework for the study of genome evolution and should help test for instance whether genome complexity originates in small effective population sizes (Lynch, 2007). Additionally, it would provide an excellent platform for genome annotation: if averaging out gene/region alignment improves robustness and accuracy for a range of estimates (Wong *et al.*, 2008; Satija *et al.*, 2008; Miklós *et al.*, 2008), one can expect that averaging out genome alignment will improve annotation accuracy, as annotation is already known to benefit from the comparative approach (Dewey *et al.*, 2004; Engelhardt *et al.*, 2005; Stanke *et al.*, 2006).

11 Phenogenomics

Beyond information on molecular evolution and phylogeny, genomes conceal the footprints of ancient functions, environmental constraints and selection pressures. Reconstructing the encoded phenotype of ancestral organisms based on the analysis of the genomes of extant individuals may be the ultimate challenge of evolutionary genomics. Recent attempts at combining models of phenotype and genome evolution foreshadow the integrative approach that could solve this problem. It is possible to infer parts of an ancient organism phenotype by reconstructing and characterizing one of its gene (for instance, (Gaucher *et al.*, 2003, 2008)) or by statistically predicting gene function with a model of function evolution (Engelhardt *et al.*, 2005). Such an approach could be extrapolated to the entire gene repertoire, thus improving previous approaches that inferred phenotype based on gene content (Mirkin *et al.*, 2003; Boussau *et al.*, 2004; Makarova *et al.*, 2006).

However, much of a phenotype emerges from a network of interactions among genes. Using approaches such as those described in preceding sections, the entire sequence of an ancestral genome may be inferred, although with a great uncertainty for weakly constrained genome parts or for very ancient organisms. From such sequences, an ancestral gene network can be resconstructed to infer the metabolic abilities of an ancient organism, or get a glimpse at its morphology.

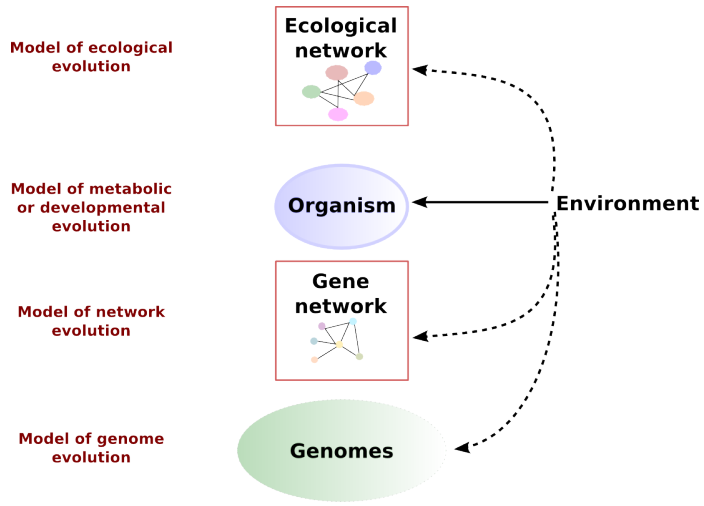


Figure 3: Models of genome evolution: from raw genome sequences to ecosystems. Organisms can be described at different levels of organisation, and for each level different models of evolution can be devised. Jointly using these models of evolution can benefit the reconstruction of characteristics of ancient organisms. The environment in which organisms live operates a direct selective pressure on the organism (continuous arrow), which has repercussions on each organisation level (dashed arrows). Genomes may adapt to the environment by shifting the composition of their genes, gene networks through the loss or acquisition of new genes and new interactions between genes, and ecological networks by the appearance or disappearance of new connections between organisms or altogether new organisms. There is thus a relation between the way organisms function and their environment, which can be input into models of evolution for more accuracy.

To reconstruct the inner workings of million-years old cells, one would have to infer the entire signalling pathways and complete metabolic cycles from genome data (Fig. 3). The reconstruction of regulation and metabolic networks has been shown to greatly benefit from an explicit evolutionary model (Wiuf *et al.*, 2006; Ratmann *et al.*, 2007; Pinney *et al.*, 2007). Pinney *et al.* (2007) used a model of network evolution to reconstruct the interaction network among bZIP transcription factors in the chordate ancestor. First, a phylogenetic tree of chordate bZIP proteins was built, and each node was identified as duplication or speciation events. This information was then crossed with interaction data from extant proteins and a model of network evolution was used to infer interactions in various ancestors in the Chordate tree. Interestingly, these authors were able to show the robustness of this approach to experimental artefacts in determining interactions between proteins. Such studies demonstrate the importance of evolutionary models to improve predictions of protein interactions and genome annotation in general, as well as ancestral network reconstruction.

Once reconstructed, ancestral networks can be linked to phenotypic characteristics. Noticeably, metabolic networks can be associated to the ecology of a microorganism, through "constraint-based reconstruction and analysis" (Price *et al.*, 2004). Using such an approach, Pál *et al.* (2006) simulated the reductive evolution of intra-cellular endosymbiotic bacteria, and assessed the relative importance of stochasticity and selection in the process. By randomly deleting genes from the genome of a free living Bacteria, monitoring the resulting impact on the metabolic network and constantly selecting for viable intermediates these authors were able to simulate a large number of artificial endosymbiont genomes. The resulting artificial genomes were similar to each other and to actual endosymbiont genomes, which reflected selection for viable cells. However, they also showed interesting differences inherent to the random order of gene deletions. In addition to simulations, "constraint-based reconstruction and analysis" could be used for reconstruction: each reconstructed ancestral

genome could be estimated under the constraint that the metabolic network it encodes should be able to sustain a cell.

Signalling networks may also be linked to phenotypic variables, so that both could be jointly reconstructed. For instance, based on empirical studies in mouse, Kavanagh *et al.* (2007) built a model of tooth development which predicts the number and type of mollars based on the relative diffusion of two types of proteins in the jaw mesenchyme. The model, which can predict dentition patterns in murine rodents, might be able to predict tooth development among all mammals (Polly, 2007). The signalling pathway ruling the development of teeth thus seems to have been broadly conserved for tens of million years. However, even for networks that have undergone profound rearrangements in evolution, it may be possible to reconstruct ancestral signalling networks jointly with phenotypic quantities using models of evolution of a continuous or discrete character (Pagel, 1999; Pagel *et al.*, 2004).

12 Ecologenomics

Among the scars that evolution has engraved in genomes are the traces of earth ancient ecosystems and climates, geographical obstacles and geological catastrophes. Quite remarkably, models of sequence evolution can also help assess ancient environments, transitory events and interactions among organisms.

For instance, mass extinctions may have left traces in the diversification pattern of species by promoting the development of new branches of the tree of life on the ashes of the disaster. Testing this prediction in mammals, Bininda-Emonds *et al.* (2007) found that the end-cretaceous extinction of dinosaurs did not increase their diversification rate. However, broader genomes samples, better models of tree reconstruction combined with models of diversification and extinction should improve our understanding of the factors influencing speciation and extinction rates (Ricklefs, 2007).

Similarly, by distinguishing species tree and gene trees, models of trans-specific polymorphisms can detect if a population is in the process of speciation (Knowles et Carstens, 2007). Studying speciation in this framework could allow coupling external variables to assess the role of geographical or ecological factors in the process. Integrative studies combining geography, ecology and sequence evolution are currently appearing, and carry much promise (Wiens *et al.*, 2007; Kozak *et al.*, 2008).

One other type of external variable that could affect the evolution of a species is found in other species with which it interacts. For instance, it has been hypothesized that the rise of angiosperms triggered the diversification of ants (Wilson et Hölldobler, 2005; Moreau *et al.*, 2006). Generally, co-evolution between species has been pervasive in the history of life, from examples as intricate as endosymbioses of mitochondrion or chloroplast to co-evolution between a host and its parasite (Biek *et al.*, 2006; Linz *et al.*, 2007). As the evolution of two interacting partners informs each other's, explicitly modelling such interactions would benefit the inference of trees and life traits. Models of ecological evolution could be superimposed on models of genome evolution to illuminate ancestral ecosystems through the reconstruction of ancestral ecological networks (Fig. 3). Recently, Dunne *et al.* (2008) showed that food webs inferred from well-preserved fossil assemblages dating from the Cambrian (>500 million years ago) were very similar to extant ones. Timeless rules thus constrain ecological networks, suggesting that models of evolution could be devised to reconstruct realistic ancestral food webs. From the study of coevolving species, one could go to the study of a whole network of interacting organisms, and infer how they evolved through time, what were the changes in the interacting partners or when a given species entered or left the network. Fully integrating evolution in the study of ecology may prove useful if we are to predict how communities may react when affected by disturbance in trophic webs or by abrupt climate changes.

13 Geogenomics

If some species get extinct when confronted with a drastic environmental change, others survive and evolve ways to cope with a new environment: a crisis may not be detectable only by its footprint on the pattern of diversification and extinction, but because genomes have kept traces of an adaptative response. For instance Christin *et al.* (2008) tested the hypothesis that a drop in atmospheric CO_2 32-25 million years ago triggered the development of C_4 metabolism in grasses, as C_4 metabolism is more efficient than the more common

C_3 metabolism. To this end, they built a large dated phylogeny of extant grass species, and estimated which ancestors already had a C_4 metabolism through a probabilistic model. Then they tested whether C_4 metabolism had appeared significantly more often after the drop in atmospheric CO_2 than before. Their results showed that after 27.6 million years, transitions to C_4 metabolism occured at a very high rate and transitions to C_3 metabolism at a nearly null rate, whereas before that date, the rate of transitions to C_4 metabolism was null: this confirms there is a correlation between atmospheric CO_2 and the evolution of C_4 metabolism.

Prokaryotic organisms adapt to temperature through changes in the nucleotidic content of their ribosomal RNAs (rRNAs) (Galtier et Lobry, 1997) as well the amino-acid content of their proteins (Zeldovich *et al.*, 2007). Consequently, by reconstructing ancient gene sequences, optimal growth temperature of extinct organisms can be inferred. (Galtier *et al.*, 1999) reconstructed ancient rRNA compositions of the last universal common ancestor (LUCA) and found support for its mesophily. More recently, Boussau *et al.* (2008) estimated growth temperatures for all prokaryotes in the tree of life, using both rRNAs and protein sequences, and found evidence for two phases in the history of environmental temperatures: thermotolerance first increased from a mesophilic LUCA to thermophilic ancestors of Bacteria and Archaea; and then steadily decreased in the bacterial kingdom. This second phase had been previously reported by another study relying on ancient gene resurrection (Gaucher *et al.*, 2008), that showed that this decrease was very similar to the evolution of ocean temperatures over the last 3.5 billion years (Robert et Chaussidon, 2006). Bacteria as a whole may thus have continuously adapted their optimal growth temperatures to the average temperature on Earth. Although this hypothesis is appealing and seems to fit the data very well, it would be even more convincing if it relied on a statistical test of the correlation between these two tendencies. To this end, a model could be built that reconstructs ancient gene composition, and simultaneously assesses the correlation with environmental temperature as inferred from geology (Lartillot, 2008).

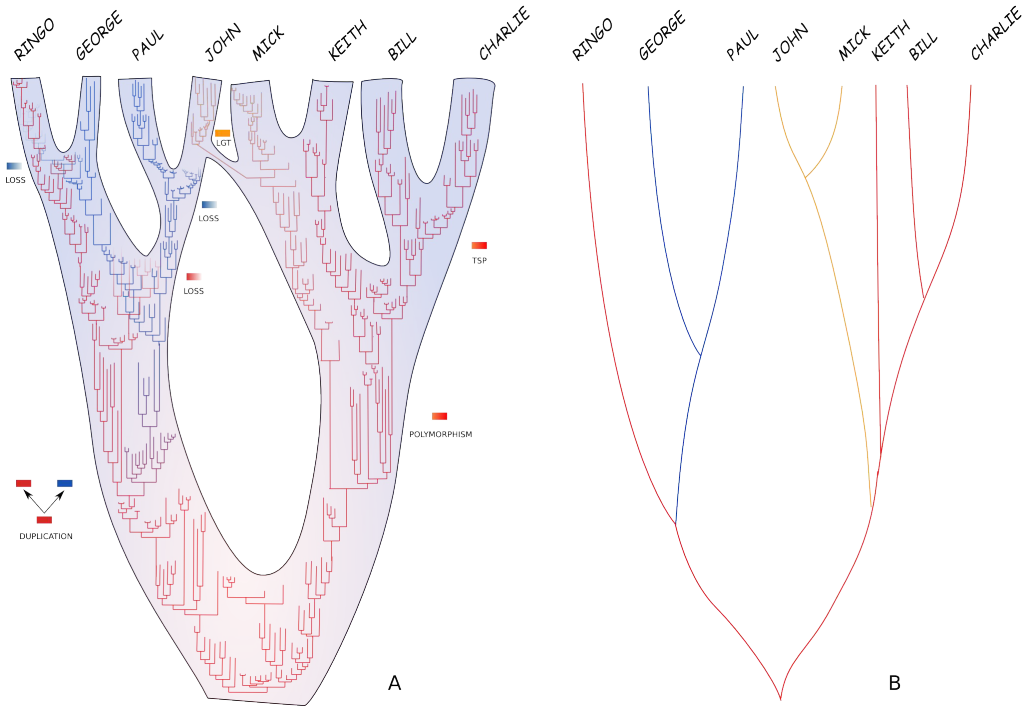
Overall, modelling interactions between species and with their environment through spatial or ecological variables could greatly benefit the study of evolution. This should enable statistical tests of the significance of correlations between environmental conditions and biological phenomena, and will enlighten how Earth and its biosphere shaped each other in their billions of years of co-existence.

14 Conclusions and perspectives

"The past is never dead. It's not even past" (Faulkner, 1951). The chronicles of life resonates in extant genomes and we have only started to exploit the historical potential of macromolecules. New statistical models of evolution, that fully exploit substitutions, gene duplication, loss and transfer, insertion-deletions and rearrangements may not only yield a better resolved tree of life, but also a thorough representation of the whole process. Through correlations between genomic properties and non-genomic variables, genomes further document the history of interactions of organisms with each other and their environment. We can now envision the reconstruction of the history of genomes along with those of metabolic or signalling networks, phenotypes and environments.

Boxes

Box 1: Discordance between a species tree and gene trees



The identification of orthologous genes is not always unequivocal. First, Phylogeneticists usually rely on the absence of duplicated copies in the datasets under study, but duplications may have occurred during the history of a gene family without leaving obvious traces. This is particularly dramatic in the event of reciprocal losses, when two species lose different copies of an ancestrally duplicated gene. The impact of this phenomenon, known as *hidden paralogy*, is difficult to estimate on a large scale, but reciprocal losses have been shown to be frequent after whole genome duplications in yeasts and fishes (Sémon et Wolfe, 2007b; Scannell *et al.*, 2007). Second, *Lateral gene transfer* (LGT) has been shown to be pervasive in the history of life, and it is unsafe to assume *a priori* that the history of a gene is devoid of such events, whatever its function. Third, even genes that would be considered genuine orthologs may not retrace the history of species: the persistence of different allelic forms of a gene during long periods of time relative to the lapse between speciation events, a phenomenon known as TRANS-SPECIFIC POLYMORPHISMS (TSP) (Wiuf *et al.*, 2004), may result in differences among gene trees (*incomplete lineage sorting* (ILS)) even in the absence of paralogy or LGT. The assemblage of these processes makes it difficult to expect that a single gene history would faithfully mirror a tree of species throughout several billion years of evolution. In addition to these biological problems, even the most advanced phylogenetic methods are often unable to accurately model the evolution of biological sequences, which can result in the inference of erroneous trees.

Box 2: Chosing the right numbers of parameters

Many processes have significantly contributed to the evolution of genomes, some related to the internal mechanics of the cell, others to the interaction of the organism with its environment and other organisms. It may be unreasonnable to devise a model accounting for all processes at once: only a limited quantity of parameters can be estimated from a finite amount of data (Steel, 2005). Bayesian MCMC techniques can certainly tolerate higher dimensionalities than Maximum Likelihood (ML) approaches, because MCMC integrates over the distributions of parameters when ML only uses point estimates: in this latter case,

any small error over the value of a variable can have snow-ball effects over the accuracy of other estimates, especially when a large number of parameters are used. However, most certainly choices will have to be made, and different models will be used depending on the question under scrutiny, where only the most relevant parameters are included: too few parameters and estimates are biased; too many, and their large variances prohibit any conclusion. In this respect, issues of model selection will be particularly pressing. Recently, several works estimated the values of parameters, but also their numbers, through techniques such as Dirichlet Process priors (Lartillot et Philippe, 2004; Huelsenbeck *et al.*, 2006), reversible jump Markov Chain Monte Carlo methods (Suchard *et al.*, 2001; Huelsenbeck *et al.*, 2004), or Poisson processes (Huelsenbeck *et al.*, 2000; Blanquart et Lartillot, 2006). Such techniques which auto-regulate their number of parameters will be necessary to use complex models with large amounts of data.

Box 3: Soft-aware

Species and gene trees

- Best (Bayesian Estimation of Species Tree): Bayesian program to reconstruct species trees from gene alignments accounting for Trans-Specific polymorphisms. http://www.stat.osu.edu/ dkp/BEST/
- Bucky (Bayesian Untangling of Concordance Knots): Bayesian program permitting to analyse several gene families simultaneously, accounting for some correlations between gene histories through gene-to-trees maps. http://www.stat.wisc.edu/ larget/bucky.html
- Prime: Set of software to analyse gene families in the presence of duplications and losses accounting for a known species tree. http://prime.sbc.su.se/

Alignment and phylogeny

- BAli-Phy (Bayesian Alignment and Phylogeny estimation): Bayesian program to reconstruct alignments and phylogenetic trees. http://www.biomath.ucla.edu/msuchard/bali-phy/index.php
- StatAlign: Bayesian program to reconstruct alignments and phylogenetic trees. http://phylogeny-cafe.elte.hu/StatAlign/
- SimulFold: Bayesian program to reconstruct RNA structural alignment as well as phylogenetic trees. http://www.cs.ubc.ca/ irmtraud/simulfold/
- Dart (DNA, Amino and RNA Tests): Software package to build and analyse alignments and phylogenetic trees through transducers notably, for sequences as well as RNA secondary structures. http://biowiki.org/DART

Inversions and phylogeny

- Badger (Bayesian Analysis to Describe Genomic Evolution by Rearrangement): Badger is a Bayesian program to analyse genomic evolution through inversions. http://badger.duq.edu/

Character evolution

- Sifter (Statistical Inference of Function Through Evolutionary Relationships): Sifter predicts the function of genes in a gene family based on a model of function evolution and on a phylogenetic tree of the gene family. http://sifter.berkeley.edu/
- Mesquite: Mesquite is a modular software gathering several packages allowing to run various types of analyses. It allows one to analyse the evolution of discrete or continuous characters on a phylogeny as well as the shape of a phylogeny. http://mesquiteproject.org/mesquite/mesquite.html
- BayesTraits: Bayesian program allowing one to analyse the evolution of discrete or continuous characters on a distribution of phylogenies. http://www.evolution.reading.ac.uk/BayesTraits.html

- Ape: Package of functions to use in the R statistical software. Ape notably permits analysing the evolution of discrete or continuous characters on a phylogeny, or studying shapes of phylogenies. http://ape.mpl.ird.fr/

Definitions

Trans-specific polymorphisms (TSPs) The sharing among species of alleles inherited from an ancestor. These alleles have diverged prior to speciation, so that gene trees reconstructed using these genes may be different from the species tree.

Incomplete lineage sorting (ILS) Observed discrepancy between a gene tree and the species tree, due to the conservation of ancestral polymorphisms in different species (trans-specific polymorphisms).

Markov Model Probabilistic model of a process in which the state at time $t+1$ only depends on state at time t, not at time $t-1$. Models of substitution assume the substitution process is markovian: a substitution $x \rightarrow y$ does not depend on the state preceding x.

Hidden Markov Model (HMM) Probabilistic model used to describe a succession of states by associating hidden states with observed ones; a Markov Model is used to describe transitions between these hidden states. Such hidden states may be "intron", "intergenic" or "exon" for models predicting gene structure, "slow" or "fast" for models predicting evolutionary rate, or different tree topologies for models predicting gene trees or recombination.

Maximum Likelihood inference (ML) For a given probabilistic model M with specific parameters and particular data D, the Maximum Likelihood values of these parameters correspond to the values under which it is most probable that the model has generated the data. If one was to simulate new data with model M, this is using these ML values that data D would be obtained most often.

Bayesian inference Likelihood is the probability of the data given the model; Bayesian inference instead deals with the probability of the model given the data, also named "posterior probability". This posterior probability of a model is proportional to the product of the likelihood and of a "prior probability". Such a prior probability permits to incorporate exterior knowledge into an analysis: for instance, one could assume that the prior probability over the transition/transversion ratio in a particular dataset follows a uniform distribution on $[1; 10]$. Contrary to Maximum Likelihood inference, the common practice in Bayesian inference is not to return parameter values of highest posterior probability; instead, whole distributions of parameter values are returned. To obtain these distributions, MCMC techniques are often used.

Markov Chain Monte Carlo (MCMC) Algorithm used to sample from a probability distribution, by building a Markov model whose equilibrium distribution is the desired probability distribution. This means that when the chain has been run for a sufficiently long time, each state is visited with a frequency equal to its probability.

Last Universal Common Ancestor (LUCA) The most recent ancestor of Bacteria, Archaea and Eukaryotes.

References

Addario-Berry, Louigi, Hallett, Michael T., et Lagergren, Jens. 2003. Towards Identifying Lateral Gene Transfer Events. *Pages 279–290 of: Pacific Symposium on Biocomputing.* 7

Ané, Cécile, Larget, Bret, Baum, David A, Smith, Stacey D, et Rokas, Antonis. 2007. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, **24**(2), 412–426. 8

Arvestad, Lars, Berglund, Ann-Charlotte, Lagergren, Jens, et Sennblad, Bengt. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, **19 Suppl 1**, i7–15. 6

Arvestad, Lars, Berglund, Ann-Charlotte, Lagergren, Jens, et Sennblad, Bengt. 2004. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution,. *In: RECOMB04.* 6

Aury, Jean-Marc, Jaillon, Olivier, Duret, Laurent, Noel, Benjamin, Jubin, Claire, Porcel, Betina M, Ségurens, Béatrice, Daubin, Vincent, Anthouard, Véronique, Aiach, Nathalie, Arnaiz, Olivier, Billaut, Alain, Beisson, Janine, Blanc, Isabelle, Bouhouche, Khaled, Câmara, Francisco, Duharcourt, Sandra, Guigo, Roderic, Gogendeau, Delphine, Katinka, Michael, Keller, Anne-Marie, Kissmehl, Roland, Klotz, Catherine, Koll, France, Mouël, Anne Le, Lepère, Gersende, Malinsky, Sophie, Nowacki, Mariusz, Nowak, Jacek K, Plattner, Helmut, Poulain, Julie, Ruiz, Françoise, Serrano, Vincent, Zagulski, Marek, Dessen, Philippe, Bétermier, Mireille, Weissenbach, Jean, Scarpelli, Claude, Schächter, Vincent, Sperling, Linda, Meyer, Eric, Cohen, Jean, et Wincker, Patrick. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, **444**(7116), 171–178. 8

Bapteste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R. L., et Doolittle, W. F. 2005. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol*, **5**(1), 33. 2

Barker, Daniel, et Pagel, Mark. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol*, **1**(1), e3. 8

Battistuzzi, Fabia U, Feijao, Andreia, et Hedges, S. Blair. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol*, **4**(Nov), 44. 2

Beerli, P., et Felsenstein, J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**(2), 763–773. 4, 5

Beiko, Robert G, Harlow, Timothy J, et Ragan, Mark A. 2005. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A*, **102**(40), 14332–14337. 7

Biek, Roman, Drummond, Alexei J, et Poss, Mary. 2006. A virus reveals population structure and recent demographic history of its carnivore host. *Science*, **311**(5760), 538–541. 12

Bininda-Emonds, Olaf R P, Cardillo, Marcel, Jones, Kate E, MacPhee, Ross D E, Beck, Robin M D, Grenyer, Richard, Price, Samantha A, Vos, Rutger A, Gittleman, John L, et Purvis, Andy. 2007. The delayed rise of present-day mammals. *Nature*, **446**(7135), 507–512. 12

Bishop, M. J., et Thompson, E. A. 1986. Maximum likelihood alignment of DNA sequences. *J Mol Biol*, **190**(2), 159–165. 8

Blanquart, Samuel, et Lartillot, Nicolas. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol Biol Evol*, **23**(11), 2058–2071. 15

Boussau, Bastien, Karlberg, E. Olof, Frank, A. Carolin, Legault, Boris-Antoine, et Andersson, Siv G E. 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc Natl Acad Sci U S A*, **101**(26), 9722–9727. 8, 10

Boussau, Bastien, Blanquart, Samuel, Necsulea, Anamaria, Lartillot, Nicolas, et Gouy., Manolo. 2008. Parallel Adaptations to High Temperatures in the Archean Eon. *Nature*. 13

Bradley, Robert K, et Holmes, Ian. 2007. Transducers: an emerging probabilistic framework for modeling indels on trees. *Bioinformatics*, **23**(23), 3258–3262. 8

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., et Stanhope, M. J. 2001. Universal trees based on large combined protein sequence data sets. *Nat Genet*, **28**(3), 281–285. 2, 3

Burt, D. W., Bruley, C., Dunn, I. C., Jones, C. T., Ramage, A., Law, A. S., Morrice, D. R., Paton, I. R., Smith, J., Windsor, D., Sazanov, A., Fries, R., et Waddington, D. 1999. The dynamics of chromosome evolution in birds and mammals. *Nature*, **402**(6760), 411–413. 10

Carstens, Bryan C, et Knowles, L. Lacey. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. *Syst Biol*, **56**(3), 400–411. 4, 5

Choi, In-Geol, et Kim, Sung-Hou. 2007. Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A*, **104**(11), 4489–4494. 7

Christin, Pascal-Antoine, Besnard, Guillaume, Samaritani, Emanuela, Duvall, Melvin R, Hodkinson, Trevor R, Savolainen, Vincent, et Salamin, Nicolas. 2008. Oligocene CO2 decline promoted C4 photosynthesis in grasses. *Curr Biol*, **18**(1), 37–43. 12

Ciccarelli, Francesca D, Doerks, Tobias, von Mering, Christian, Creevey, Christopher J, Snel, Berend, et Bork, Peer. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**(5765), 1283–1287. 2

Dagan, Tal, et Martin, William. 2006. The tree of one percent. *Genome Biol*, **7**(10), 118. 2

Dagan, Tal, et Martin, William. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci U S A*, **104**(3), 870–875. 7

Darling, Aaron E, Miklós, István, et Ragan, Mark A. 2008. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet*, **4**(7), e1000128. 10

Darwin, Charles R. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* 1st edn. London: John Murray. 2

Daubin, Vincent, et Perrière, Guy. 2003. G+C3 structuring along the genome: a common feature in prokaryotes. *Mol Biol Evol*, **20**(4), 471–483. 10

Daubin, Vincent, Moran, Nancy A, et Ochman, Howard. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science*, **301**(5634), 829–832. 7

Degnan, James H, et Rosenberg, Noah A. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet*, **2**(5), e68. 3, 4

Delsuc, Frédéric, Brinkmann, Henner, et Philippe, Hervé. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, **6**(5), 361–375. 2

Dewey, Colin, Wu, Jia Qian, Cawley, Simon, Alexandersson, Marina, Gibbs, Richard, et Pachter, Lior. 2004. Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Res*, **14**(4), 661–664. 10

Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. *Science*, **284**(5423), 2124–2129. 2, 7

Dubb, Lindsey. 2005. *A Likelihood Model of Gene Family Evolution.* Ph.D. thesis, University of Washington. 6

Dunn, Casey W, Hejnol, Andreas, Matus, David Q, Pang, Kevin, Browne, William E, Smith, Stephen A, Seaver, Elaine, Rouse, Greg W, Obst, Matthias, Edgecombe, Gregory D, Sørensen, Martin V, Haddock, Steven H D, Schmidt-Rhaesa, Andreas, Okusu, Akiko, Kristensen, Reinhardt Møbjerg, Wheeler, Ward C, Martindale, Mark Q, et Giribet, Gonzalo. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**(7188), 745–749. 6

Dunne, Jennifer A, Williams, Richard J, Martinez, Neo D, Wood, Rachel A, et Erwin, Douglas H. 2008. Compilation and network analyses of cambrian food webs. *PLoS Biol*, **6**(4), e102. 12

Duret, L., Dorkeld, F., et Gautier, C. 1993. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res*, **21**(10), 2315–2322. 9

Edgar, Robert C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**(Aug), 113. 8

Engelhardt, Barbara E, Jordan, Michael I, Muratore, Kathryn E, et Brenner, Steven E. 2005. Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput Biol*, **1**(5), e45. 10

Eyre-Walker, A., et Hurst, L. D. 2001. The evolution of isochores. *Nat Rev Genet*, **2**(7), 549–555. 10

Faulkner, William. 1951. *Requiem for a Nun*. Penguin books. 13

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool*, **19**(2), 99–113. 2

Galtier, N., et Lobry, J. R. 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*, **44**(6), 632–636. 13

Galtier, N., Tourasse, N., et Gouy, M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science*, **283**(5399), 220–221. 13

Gaucher, E. A., Thomson, J. M., Burgan, M. F., et Benner, S. A. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature*, **425**(6955), 285–288. 10

Gaucher, Eric A, Govindarajan, Sridhar, et Ganesh, Omjoy K. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature*, **451**(7179), 704–707. 10, 13

Ge, Fan, Wang, Li-San, et Kim, Junhyong. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. *PLoS Biol*, **3**(10), e316. 7

Hallett, M.T., Lagergren, L., et Tofigh, A. 2004. Simulateous Identification of Gene Duplication and Horizontal Transfer Events. *In: RECOMB*. 7

Hao, Weilong, et Golding, G. Brian. 2008. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics*, **9**, 235. 8

Hein, J. 2001. An algorithm for statistical alignment of sequences related by a binary tree. *Pac Symp Biocomput*, 179–190. 8

Hobolth, Asger, Christensen, Ole F, Mailund, Thomas, et Schierup, Mikkel H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet*, **3**(2), e7. 8

Holmes, I., et Bruno, W. J. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, **17**(9), 803–820. 8

Huang, Jinling, et Gogarten, Johann Peter. 2006. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. *Trends Genet*, **22**(7), 361–366. 7

Huelsenbeck, J. P., et Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**(8), 754–755. 6

Huelsenbeck, J. P., Larget, B., et Swofford, D. 2000. A compound poisson process for relaxing the molecular clock. *Genetics*, **154**(4), 1879–1892. 15

Huelsenbeck, John P, Larget, Bret, Miller, Richard E, et Ronquist, Fredrik. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst Biol*, **51**(5), 673–688. 9

Huelsenbeck, John P, Larget, Bret, et Alfaro, Michael E. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol Biol Evol*, **21**(6), 1123–1133. 15

Huelsenbeck, John P, Jain, Sonia, Frost, Simon W D, et Pond, Sergei L Kosakovsky. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci U S A*, **103**(16), 6263–6268. 15

Innan, Hideki, et Watanabe, Hidemi. 2006. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Mol Biol Evol*, **23**(5), 1040–1047. 6

Kavanagh, Kathryn D, Evans, Alistair R, et Jernvall, Jukka. 2007. Predicting evolutionary patterns of mammalian teeth from development. *Nature*, **449**(7161), 427–432. 12

Kedzierska, Anna, et Husmeier, Dirk. 2006. A heuristic Bayesian method for segmenting DNA sequence alignments and detecting evidence for recombination and gene conversion. *Stat Appl Genet Mol Biol*, **5**, Article27. 9

Kingman, JFC. 1982. On the genealogy of large populations. *Journal of Applied Probability*, **19A**, 27–43. 4

Knowles, L. Lacey, et Carstens, Bryan C. 2007. Delimiting species without monophyletic gene trees. *Syst Biol*, **56**(6), 887–895. 12

Kozak, Kenneth H, Graham, Catherine H, et Wiens, John J. 2008. Integrating GIS-based environmental data into evolutionary biology. *Trends Ecol Evol*, **23**(3), 141–148. 12

Kubatko, Laura Salter, et Degnan, James H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*, **56**(1), 17–24. 4

Kurland, C. G., Canback, B., et Berg, Otto G. 2003. Horizontal gene transfer: a critical view. *Proc Natl Acad Sci U S A*, **100**(17), 9658–9662. 7

Lake, J. A. 1991. The order of sequence alignment can bias the selection of tree topology. *Mol Biol Evol*, **8**(3), 378–385. 8

Landan, Giddy, et Graur, Dan. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. *Mol Biol Evol*, **24**(6), 1380–1383. 8

Larget, Bret, Kadane, Joseph B, et Simon, Donald L. 2005. A Bayesian approach to the estimation of ancestral genome arrangements. *Mol Phylogenet Evol*, **36**(2), 214–223. 10

Lartillot, Nicolas. 2008. *Personal communication.* 13

Lartillot, Nicolas, et Philippe, Hervé. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, **21**(6), 1095–1109. 15

Lerat, Emmanuelle, Daubin, Vincent, et Moran, Nancy A. 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS Biol*, **1**(1), E19. 7

Linz, Bodo, Balloux, François, Moodley, Yoshan, Manica, Andrea, Liu, Hua, Roumagnac, Philippe, Falush, Daniel, Stamer, Christiana, Prugnolle, Franck, van der Merwe, Schalk W, Yamaoka, Yoshio, Graham, David Y, Perez-Trallero, Emilio, Wadstrom, Torkel, Suerbaum, Sebastian, et Achtman, Mark. 2007. An African origin for the intimate association between humans and Helicobacter pylori. *Nature*, **445**(7130), 915–918. 12

Liu, Liang, et Pearl, Dennis K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol*, **56**(3), 504–514. 4, 5

Lobry, J. R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*, **13**(5), 660–665. 10

Long, Manyuan, Betrán, Esther, Thornton, Kevin, et Wang, Wen. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*, **4**(11), 865–875. 10

Lunter, Gerton, Miklós, István, Drummond, Alexei, Jensen, Jens Ledet, et Hein, Jotun. 2005. Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, **6**, 83. 8

Lynch, Michael. 2007. *The Origins of Genome Architecture*. Sinauer Assocs., Inc., Sunderland, MA. 10

Löytynoja, Ari, et Goldman, Nick. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**(5883), 1632–1635. 8

Maddison, Wayne P. 1997. Gene trees in species trees. *Syst Biol*, **46**, 523–536. 3

Maddison, Wayne P, et Knowles, L. Lacey. 2006. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*, **55**(1), 21–30. 4, 5

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N., Shakhova, V., Grigoriev, I., Lou, Y., Rohksar, D., Lucas, S., Huang, K., Goodstein, D. M., Hawkins, T., Plengvidhya, V., Welker, D., Hughes, J., Goh, Y., Benson, A., Baldwin, K., Lee, J-H., Díaz-Muñiz, I., Dosti, B., Smeianov, V., Wechter, W., Barabote, R., Lorca, G., Altermann, E., Barrangou, R., Ganesan, B., Xie, Y., Rawsthorne, H., Tamir, D., Parker, C., Breidt, F., Broadbent, J., Hutkins, R., O'Sullivan, D., Steele, J., Unlu, G., Saier, M., Klaenhammer, T., Richardson, P., Kozyavkin, S., Weimer, B., et Mills, D. 2006. Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A*, **103**(42), 15611–15616. 10

Metzler, Dirk. 2003. Statistical alignment based on fragment insertion and deletion models. *Bioinformatics*, **19**(4), 490–499. 8

Miklós, István, Novák, Adám, Dombai, Balázs, et Hein, Jotun. 2008. How reliably can we predict the reliability of protein structure predictions? *BMC Bioinformatics*, **9**, 137. 9, 10

Minin, Vladimir N, Dorman, Karin S, Fang, Fang, et Suchard, Marc A. 2005. Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics*, **21**(13), 3034–3042. 9

Mirkin, Boris G, Fenner, Trevor I, Galperin, Michael Y, et Koonin, Eugene V. 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**(Jan), 2. 10

Mitchison, G. J. 1999. A probabilistic treatment of phylogeny and sequence alignment. *J Mol Evol*, **49**(1), 11–22. 8

Moreau, Corrie S, Bell, Charles D, Vila, Roger, Archibald, S. Bruce, et Pierce, Naomi E. 2006. Phylogeny of the ants: diversification in the age of angiosperms. *Science*, **312**(5770), 101–104. 12

Muffato, Matthieu, et Crollius, Hugues Roest. 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. *Bioessays*, **30**(2), 122–134. 10

Necsulea, Anamaria, et Lobry, Jean R. 2007. A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol Biol Evol*, **24**(10), 2169–2179. 10

Notredame, C., Higgins, D. G., et Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–217. 8

Ochman, H., Lawrence, J. G., et Groisman, E. A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**(6784), 299–304. 7

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature*, **401**(6756), 877–884. 12

Pagel, Mark, Meade, Andrew, et Barker, Daniel. 2004. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*, **53**(5), 673–684. 12

Pinney, John W, Amoutzias, Grigoris D, Rattray, Magnus, et Robertson, David L. 2007. Reconstruction of ancestral protein interaction networks for the bZIP transcription factors. *Proc Natl Acad Sci U S A*, **104**(51), 20449–20453. 11

Polly, P. David. 2007. Evolutionary biology: development with a bite. *Nature*, **449**(7161), 413–415. 12

Pond, Sergei L Kosakovsky, Posada, David, Gravenor, Michael B, Woelk, Christopher H, et Frost, Simon D W. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol*, **23**(10), 1891–1901. 9

Price, Nathan D, Reed, Jennifer L, et Palsson, Bernhard Ø. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol*, **2**(11), 886–897. 11

Pál, Csaba, Papp, Balázs, Lercher, Martin J, Csermely, Péter, Oliver, Stephen G, et Hurst, Laurence D. 2006. Chance and necessity in the evolution of minimal metabolic networks. *Nature*, **440**(7084), 667–670. 11

Rannala, Bruce, et Yang, Ziheng. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**(4), 1645–1656. 4, 5

Ratmann, Oliver, Jørgensen, Ole, Hinkley, Trevor, Stumpf, Michael, Richardson, Sylvia, et Wiuf, Carsten. 2007. Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Comput Biol*, **3**(11), e230. 11

Redelings, Benjamin D, et Suchard, Marc A. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst Biol*, **54**(3), 401–418. 8

Redelings, Benjamin D, et Suchard, Marc A. 2007. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evol Biol*, **7**, 40. 9

Ricklefs, Robert E. 2007. Estimating diversification rates from phylogenetic information. *Trends Ecol Evol*, **22**(11), 601–610. 12

Rieseberg, L. 2001. Chromosomal rearrangements and speciation. *Trends Ecol Evol*, **16**(7), 351–358. 10

Robert, F., et Chaussidon, M. 2006. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature*, **443**(7114), 969–972. 13

Rokas, et Holland. 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol*, **15**(11), 454–459. 3

Rokas, Antonis, et Carroll, Sean B. 2006. Bushes in the tree of life. *PLoS Biol*, **4**(11), e352. 2

Rosenberg, Noah A, et Tao, Randa. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. *Syst Biol*, **57**(1), 131–140. 4

Satija, Rahul, Pachter, Lior, et Hein, Jotun. 2008. Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics*, **24**(10), 1236–1242. 9, 10

Scannell, Devin R, Frank, A. Carolin, Conant, Gavin C, Byrne, Kevin P, Woolfit, Megan, et Wolfe, Kenneth H. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A*, **104**(20), 8397–8402. 14

Snel, Berend, Bork, Peer, et Huynen, Martijn A. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, **12**(1), 17–25. 8

Snel, Berend, Huynen, Martijn A, et Dutilh, Bas E. 2005. Genome trees and the nature of genome evolution. *Annu Rev Microbiol*, **59**, 191–209. 2

Stanke, Mario, Tzvetkova, Ana, et Morgenstern, Burkhard. 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol*, **7 Suppl 1**, S11.1–S11.8. 10

Stark, Alexander, Lin, Michael F, Kheradpour, Pouya, Pedersen, Jakob S, Parts, Leopold, Carlson, Joseph W, Crosby, Madeline A, Rasmussen, Matthew D, Roy, Sushmita, Deoras, Ameya N, Ruby, J. Graham, Brennecke, Julius, curators, Harvard FlyBase, Project, Berkeley Drosophila Genome, Hodges, Emily, Hinrichs, Angie S, Caspi, Anat, Paten, Benedict, Park, Seung-Won, Han, Mira V, Maeder, Morgan L, Polansky, Benjamin J, Robson, Bryanne E, Aerts, Stein, van Helden, Jacques, Hassan, Bassem, Gilbert, Donald G, Eastman, Deborah A, Rice, Michael, Weir, Michael, Hahn, Matthew W, Park, Yongkyu, Dewey, Colin N, Pachter, Lior, Kent, W. James, Haussler, David, Lai, Eric C, Bartel, David P, Hannon, Gregory J, Kaufman, Thomas C, Eisen, Michael B, Clark, Andrew G, Smith, Douglas, Celniker, Susan E, Gelbart, William M, et Kellis, Manolis. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, **450**(7167), 219–232. 9

Steel, Mike. 2005. Should phylogenetic models be trying to "fit an elephant"? *Trends Genet*, **21**(6), 307–309. 14

Sturtevant, A. H., et Dobzhansky, T. 1936. Inversions in the Third Chromosome of Wild Races of Drosophila Pseudoobscura, and Their Use in the Study of the History of the Species. *Proc Natl Acad Sci U S A*, **22**(7), 448–450. 2, 10

Suchard, M. A., Weiss, R. E., et Sinsheimer, J. S. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol*, **18**(6), 1001–1013. 15

Suchard, Marc A. 2005. Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics*, **170**(1), 419–431. 7

Suchard, Marc A, Kitchen, Christina M R, Sinsheimer, Janet S, et Weiss, Robert E. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst Biol*, **52**(5), 649–664. 3

Sémon, Marie, et Wolfe, Kenneth H. 2007a. Consequences of genome duplication. *Curr Opin Genet Dev*, **17**(6), 505–512. 8, 10

Sémon, Marie, et Wolfe, Kenneth H. 2007b. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet*, **23**(3), 108–112. 14

Than, Cuong, Ruths, Derek, Innan, Hideki, et Nakhleh, Luay. 2007. Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions. *J Comput Biol*, **14**(4), 517–535. 7

Thompson, J. D., Higgins, D. G., et Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**(22), 4673–4680. 8

Thorne, J. L., Kishino, H., et Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol*, **33**(2), 114–124. 8

Thorne, J. L., Kishino, H., et Felsenstein, J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J Mol Evol*, **34**(1), 3–16. 8

Wiens, John J, Kuczynski, Caitlin A, Duellman, William E, et Reeder, Tod W. 2007. Loss and re-evolution of complex life cycles in marsupial frogs: does ancestral trait reconstruction mislead? *Evolution*, **61**(8), 1886–1899. 12

Wilson, Edward O, et Hölldobler, Bert. 2005. The rise of the ants: a phylogenetic and ecological explanation. *Proc Natl Acad Sci U S A*, **102**(21), 7411–7414. 12

Wiuf, Carsten, Zhao, Keyan, Innan, Hideki, et Nordborg, Magnus. 2004. The probability and chromosomal extent of trans-specific polymorphism. *Genetics*, **168**(4), 2363–2372. 14

Wiuf, Carsten, Brameier, Markus, Hagberg, Oskar, et Stumpf, Michael P H. 2006. A likelihood approach to analysis of network data. *Proc Natl Acad Sci U S A*, **103**(20), 7566–7570. 11

Wong, Karen M, Suchard, Marc A, et Huelsenbeck, John P. 2008. Alignment uncertainty and genomic analysis. *Science*, **319**(5862), 473–476. 8, 9, 10

Yap, Von Bing, et Speed, Terry. 2005. Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evol Biol*, **5**(1), 2. 9

Zeldovich, Konstantin B, Berezovsky, Igor N, et Shakhnovich, Eugene I. 2007. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol*, **3**(1), e5. 13

Zhaxybayeva, Olga, Gogarten, J. Peter, Charlebois, Robert L, Doolittle, W. Ford, et Papke, R. Thane. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res*, **16**(9), 1099–1108. 7

Zuckerkandl, E., et Pauling, L. 1965. Molecules as documents of evolutionary history. *J Theor Biol*, **8**(2), 357–366. 2

11 Conclusion

This thesis has attempted to tackle issues related to the early evolution of life by analysing the genomes of extant organisms. It relied on statistical approaches, as it takes maths to accurately read genomes.

In article 6, the statistical approach to genomics led to the conclusion that life started in warm conditions, then adapted to higher temperatures before decreasing again. These variations could be matched with hypotheses from geological studies.

Indeed, conclusions based on comparative genomics need to be confronted with knowledge coming from other disciplines. For early evolution, geology is the only relevant point of comparison. I believe new models of evolution should benefit from both fields of study, and combine models of genome evolution with models of the evolution of the Earth.

In this context, many examples of integrative phylogenomics, combining phylogeography, climatology or ecology with genome evolution may appear in the next years, and provide a better picture of the evolution of the Earth and of its inhabitants. This work gave me the desire to embark on such projects.

Bibliography

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B., et de Rosa, R. 2000. The new animal phylogeny: reliability and implications. *Proc. Natl. Acad. Sci. U. S. A.*, **97**(9), 4453–4456. 48

Allwood, Abigail C, Walter, Malcolm R, Kamber, Balz S, Marshall, Craig P, et Burch, Ian W. 2006. Stromatolite reef from the Early Archaean era of Australia. *Nature*, **441**(7094), 714–718. 20

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402. 26

Ané, Cécile, Larget, Bret, Baum, David A, Smith, Stacey D, et Rokas, Antonis. 2007. Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, **24**(2), 412–426. 39

AToL. 2008 (August). *Assembling the Tree of Life*. `http://atol.sdsc.edu/`. 17

Bapteste, E., Susko, E., Leigh, J., Ruiz-Trillo, I., Bucknam, J., et Doolittle, W. F. 2008. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. *Mol Biol Evol*, **25**(1), 83–91. 47

Battistuzzi, Fabia U, Feijao, Andreia, et Hedges, S. Blair. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol Biol*, **4**(Nov), 44. 47

Beiko, Robert G, Harlow, Timothy J, et Ragan, Mark A. 2005. Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci U S A*, **102**(40), 14332–14337. 47

Bekker, A., Holland, H. D., Wang, P-L., Rumble, D., Stein, H. J., Hannah, J. L., Coetzee, L. L., et Beukes, N. J. 2004. Dating the rise of atmospheric oxygen. *Nature*, **427**(6970), 117–120. 22

Bern, Marshall, et Goldberg, David. 2005. Automatic selection of representative proteins for bacterial phylogeny. *BMC Evol Biol*, **5**(1), 34. 47

Berner, Petsch, Lake, Beerling, Popp, Lane, Laws, Westley, Cassar, Woodward, et Quick. 2000. Isotope fractionation and atmospheric oxygen: implications for phanerozoic O(2) evolution. *Science*, **287**(5458), 1630–1633. 23

Bonen, L., et Doolittle, W. F. 1975. On the prokaryotic nature of red algal chloroplasts. *Proc Natl Acad Sci U S A*, **72**(6), 2310–2314. 6, 45

Bonen, L., Cunningham, R. S., Gray, M. W., et Doolittle, W. F. 1977. Wheat embryo mitochondrial 18S ribosomal RNA: evidence for its prokaryotic nature. *Nucleic Acids Res*, **4**(3), 663–671. 6, 45

Bourlat, Sarah J, Juliusdottir, Thorhildur, Lowe, Christopher J, Freeman, Robert, Aronowicz, Jochanan, Kirschner, Mark, Lander, Eric S, Thorndyke, Michael, Nakano, Hiroaki, Kohn, Andrea B, Heyland, Andreas, Moroz, Leonid L, Copley, Richard R, et Telford, Maximilian J. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature*, **444**(7115), 85–88. 48

Boussau, Bastien, et Gouy, Manolo. 2006. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol*, **55**(5), 756–768. 42

Boussau, Bastien, Karlberg, E. Olof, Frank, A. Carolin, Legault, Boris-Antoine, et Andersson, Siv G E. 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc Natl Acad Sci U S A*, **101**(26), 9722–9727. 47

Braddy, Simon J, Poschmann, Markus, et Tetlie, O. Erik. 2008. Giant claw reveals the largest ever arthropod. *Biol Lett*, **4**(1), 106–109. 23

Brochier, Celine, Gribaldo, Simonetta, Zivanovic, Yvan, Confalonieri, Fabrice, et Forterre, Patrick. 2005a. Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales? *Genome Biol*, **6**(5), R42. 46

Brochier, Céline, et Philippe, Hervé. 2002. Phylogeny: a non-hyperthermophilic ancestor for bacteria. *Nature*, **417**(6886), 244. 47

Brochier, Céline, Forterre, Patrick, et Gribaldo, Simonetta. 2004. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. *Genome Biol*, **5**(3), R17. 46

Brochier, Céline, Forterre, Patrick, et Gribaldo, Simonetta. 2005b. 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. 46

Brochier-Armanet, Céline, Boussau, Bastien, Gribaldo, Simonetta, et Forterre, Patrick. 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol*, **6**(3), 245–252. 46, 49

Brocks, J. J., Logan, G. A., Buick, R., et Summons, R. E. 1999. Archean molecular fossils and the early rise of eukaryotes. *Science*, **285**(5430), 1033–1036. 20

Brocks, Jochen J, Love, Gordon D, Summons, Roger E, Knoll, Andrew H, Logan, Graham A, et Bowden, Stephen A. 2005. Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature*, **437**(7060), 866–870. 21

Burki, Fabien, Shalchian-Tabrizi, Kamran, et Pawlowski, Jan. 2008. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. *Biol Lett*, **4**(4), 366–369. 48

Canfield, D. E., Habicht, K. S., et Thamdrup, B. 2000. The Archean sulfur cycle and the early history of atmospheric oxygen. *Science*, **288**(5466), 658–661. 22

Cavalli-Sforza, L. L., et Edwards, A. W. 1967. Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet*, **19**(3 Pt 1), 233–257. 35

Choi, In-Geol, et Kim, Sung-Hou. 2007. Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A*, **104**(11), 4489–4494. 47

Ciccarelli, Francesca D, Doerks, Tobias, von Mering, Christian, Creevey, Christopher J, Snel, Berend, et Bork, Peer. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**(5765), 1283–1287. 47

Connelly, James N., Amelin, Yuri, Krot, Alexander N., et Bizzarro, Martin. 2008. Chronology of the Solar System's Oldest Solids. *The Astrophysical Journal Letters*, **675**, L121–L124. 19

Darwin, Charles R. 1845. *Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N.* 2d edn. London: John Murray. 13

Darwin, Charles R. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* 1st edn. London: John Murray. 14

Daubin, Vincent, Moran, Nancy A, et Ochman, Howard. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science*, **301**(5634), 829–832. 47

Davies, T. Jonathan, Barraclough, Timothy G, Chase, Mark W, Soltis, Pamela S, Soltis, Douglas E, et Savolainen, Vincent. 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proc Natl Acad Sci U S A*, **101**(7), 1904–1909. 48

Dawkins, Richard. 2005. *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution.* Mariner Books. 25

Dayhoff, M.O., Eyck, R.V., et Park, C.M. 1972. *Atlas of protein sequence and structure.* Washington, DC: National Biomedical Research Foundation. Chap. A model of evolutionary change in proteins. In: Atlas of protein sequence and structure., page 89–99. 33

Delsuc, Frédéric, Brinkmann, Henner, Chourrout, Daniel, et Philippe, Hervé. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, **439**(7079), 965–968. 48

Deusch, Oliver, Landan, Giddy, Roettger, Mayo, Gruenheit, Nicole, Kowallik, Klaus V, Allen, John F, Martin, William, et Dagan, Tal. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol*, **25**(4), 748–761. 6, 45

Dobzhansky, Theodosius. 1973. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher*, **35**(March), 125–129. 15

Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. *Science*, **284**(5423), 2124–2129. 47

Douzery, Emmanuel J P, Snell, Elizabeth A, Bapteste, Eric, Delsuc, Frédéric, et Philippe, Hervé. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci U S A*, **101**(43), 15386–15391. 48

Drummond, Alexei J, Ho, Simon Y W, Phillips, Matthew J, et Rambaut, Andrew. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, **4**(5), e88. 42

Dudley, R. 1998. Atmospheric oxygen, giant Paleozoic insects and the evolution of aerial locomotor performance. *J Exp Biol*, **201**(Pt 8), 1043–1050. 23

Dunn, Casey W, Hejnol, Andreas, Matus, David Q, Pang, Kevin, Browne, William E, Smith, Stephen A, Seaver, Elaine, Rouse, Greg W, Obst, Matthias, Edgecombe, Gregory D, Sørensen, Martin V, Haddock, Steven H D, Schmidt-Rhaesa, Andreas, Okusu, Akiko, Kristensen, Reinhardt Møbjerg, Wheeler, Ward C, Martindale, Mark Q, et Giribet, Gonzalo. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**(7188), 745–749. 48

Dutheil, Julien, Gaillard, Sylvain, Bazin, Eric, Glémin, Sylvain, Ranwez, Vincent, Galtier, Nicolas, et Belkhir, Khalid. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, **7**, 188. 103

Edgar, Robert C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**(Aug), 113. 26

Edwards, A. W. F. 1972. *Likelihood*. Cambridge University Press. 39

Edwards, A.W.F., et Cavalli-Sforza, L.L. 1964. *Phenetic and Phylogenetic Classification*. Systematics Association Publisher, London. Chap. Reconstructon of evolutionary trees, pages 67–76. 35

Eiler, John M. 2007. Geochemistry. The oldest fossil or just another rock? *Science*, **317**(5841), 1046–1047. 21

Elkins, James G, Podar, Mircea, Graham, David E, Makarova, Kira S, Wolf, Yuri, Randau, Lennart, Hedlund, Brian P, Brochier-Armanet, Céline, Kunin, Victor, Anderson, Iain, Lapidus, Alla, Goltsman, Eugene, Barry, Kerrie, Koonin, Eugene V, Hugenholtz, Phil, Kyrpides, Nikos, Wanner, Gerhard, Richardson, Paul, Keller, Martin, et Stetter, Karl O. 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. *Proc Natl Acad Sci U S A*, **105**(23), 8102–8107. 46, 49

Esser, Christian, Ahmadinejad, Nahal, Wiegand, Christian, Rotte, Carmen, Sebastiani, Federico, Gelius-Dietrich, Gabriel, Henze, Katrin, Kretschmann, Ernst, Richly, Erik, Leister, Dario, Bryant, David, Steel, Michael A, Lockhart, Peter J, Penny, David, et Martin, William. 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol*, **21**(9), 1643–1660. 6, 45

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet*, **25**(5), 471–492. 33

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, **17**(6), 368–376. 33, 36

Felsenstein, Joe. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*, **27**, 401–410. 33

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sinauer Associates. 33

Fike, D. A., Grotzinger, J. P., Pratt, L. M., et Summons, R. E. 2006. Oxidation of the Ediacaran ocean. *Nature*, **444**(7120), 744–747. 22

Fisher, RA. 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London*, **222**, 309–368. 36

Fitch, Walter M. 1971. Toward defining thecourse of evolution: Minimum change for a specified tree topology. *Syst Zool*, **20**, 406–416. 35

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., et Merrick, J. M. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**(5223), 496–512. 40

Forterre, P. 2002. The origin of DNA genomes and DNA replication proteins. *Curr. Opin. Microbiol.*, **5**(5), 525–532. 10

Forterre, P., Benachenhou-Lahfa, N., Confalonieri, F., Duguet, M., Elie, C., et Labedan, B. 1992. The nature of the last universal ancestor and the root of the tree of life, still open questions. *Biosystems*, **28**(1-3), 15–32. 42

Forterre, Patrick. 2007. *Microbes de l'enfer*. Belin. 6, 40

Forterre, Patrick, Brochier, Celine, et Philippe, Hervé. 2002. Evolution of the Archaea. *Theor Popul Biol*, **61**(4), 409–422. 46

Fournier, G. P., et Gogarten, J. P. 2007. Signature of a primitive genetic code in ancient protein lineages. *J. Mol. Evol.*, **65**(4), 425–436. 105

Frohlich, Michael W, et Chase, Mark W. 2007. After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature*, **450**(7173), 1184–1189. 48

Galtier, N., et Gouy, M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol*, **15**(7), 871–879. 15, 42

Galtier, Nicolas. 1997. *L'approche statistique en phylogénie moléculaire : influence des compositions en bases variables*. Ph.D. thesis, Université Claude Bernard - Lyon 1. 33

Galtier, Nicolas. 2007. A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst Biol*, **56**(4), 633–642. 47

Gascuel, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*, **14**(7), 685–695. 35

Gaucher, Eric A, Govindarajan, Sridhar, et Ganesh, Omjoy K. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature*, **451**(7179), 704–707. 9

Geyer, Charles J. 1991. *Reweighting Monte Carlo Mixtures*. Tech. rept. 568. School of Satistics, University of Minnesota. 38

Gillespie, J. H. 1984. The molecular clock may be an episodic clock. *Proc Natl Acad Sci U S A*, **81**(24), 8009–8013. 42

Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., et Oshima, T. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.*, **86**(17), 6661–6665. 6, 41

Gomes, R., Levison, H. F., Tsiganis, K., et Morbidelli, A. 2005. Origin of the cataclysmic Late Heavy Bombardment period of the terrestrial planets. *Nature*, **435**(7041), 466–469. 10

Gouy, M., et Li, W. H. 1989. Molecular phylogeny of the kingdoms Animalia, Plantae, and Fungi. *Mol Biol Evol*, **6**(2), 109–122. 48

Graur, D., Duret, L., et Gouy, M. 1996. Phylogenetic position of the order Lagomorpha (rabbits, hares and allies). *Nature*, **379**(6563), 333–335. 48

Graur, D., Gouy, M., et Duret, L. 1997. Evolutionary affinities of the order Perissodactyla and the phylogenetic status of the superordinal taxa Ungulata and Altungulata. *Mol Phylogenet Evol*, **7**(2), 195–200. 48

Gribaldo, Simonetta, et Brochier-Armanet, Celine. 2006. The origin and evolution of Archaea: a state of the art. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **361**(1470), 1007–1022. 46, 47, 49

Guindon, Stéphane, et Gascuel, Olivier. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52**(5), 696–704. 33, 35

Guindon, Stéphane. 2003. *Méthodes et algorithmes pour l'approche statistique en phylogénie*. Ph.D. thesis, Université Montpellier II - Sciences et Techniques du Languedoc - U.F.R. Sciences de Montpellier. 33

Hasegawa, M., et Fujiwara, M. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol Phylogenet Evol*, **2**(1), 1–5. 33

Hasegawa, M., Kishino, H., et Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, **22**(2), 160–174. 33, 103

Hrdy, Ivan, Hirt, Robert P, Dolezal, Pavel, Bardonová, Lucie, Foster, Peter G, Tachezy, Jan, et Embley, T. Martin. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature*, **432**(7017), 618–622. 46

Huelsenbeck, J. P. 1997. Is the Felsenstein zone a fly trap? *Syst Biol*, **46**(1), 69–74. 33

Huelsenbeck, John P, Bollback, Jonathan P, et Levine, Amy M. 2002. Inferring the root of a phylogenetic tree. *Syst Biol*, **51**(1), 32–43. 42, 44

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., et Miyata, T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. U. S. A.*, **86**(23), 9355–9359. 6, 41

Jansen, Robert K, Cai, Zhengqiu, Raubeson, Linda A, Daniell, Henry, Depamphilis, Claude W, Leebens-Mack, James, Müller, Kai F, Guisinger-Bellian, Mary, Haberle, Rosemarie C, Hansen, Anne K, Chumley, Timothy W, Lee, Seung-Bum, Peery, Rhiannon, McNeal, Joel R, Kuehl, Jennifer V, et Boore, Jeffrey L. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci U S A*, **104**(49), 19369–19374. 48

Jones, D. T., Taylor, W. R., et Thornton, J. M. 1994. A mutation data matrix for transmembrane proteins. *FEBS Lett*, **339**(3), 269–275. 33

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. *Pages 21–132 of:* Munro, M.N. (ed), *Mammalian Protein Metabolism*, vol. 3. Academic Press New York. 33

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. 33

Knoll, Andrew H. 2004. *Life on a Young Planet: The First Three Billion Years of Evolution on Earth.* Princeton University Press. 20

Kuhner, M. K., et Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*, **11**(3), 459–468. 33

Kuhner, M. K., Yamato, J., et Felsenstein, J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**(4), 1421–1430. 38

Kumar, Sudhir. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet*, **6**(8), 654–662. 42

Lanave, C., Preparata, G., Saccone, C., et Serio, G. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, **20**(1), 86–93. 33

Lane, Nick. 2004. *Oxygen: The Molecule that Made the World.* Oxford University Press, USA. 20

Le, Si Quang, et Gascuel, Olivier. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol*, **25**(7), 1307–1320. 33

Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G. W., Prosser, J. I., Schuster, S. C., et Schleper, C. 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, **442**(7104), 806–809. 19

Li, S., Pearl, D.K., et Doss, H. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, **95**, 493–508. 38

Li, W. H., Gouy, M., Sharp, P. M., O'hUigin, C., et Yang, Y. W. 1990. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc Natl Acad Sci U S A*, **87**(17), 6703–6707. 48

Lipp, Julius S, Morono, Yuki, Inagaki, Fumio, et Hinrichs, Kai-Uwe. 2008. Significant contribution of Archaea to extant biomass in marine subsurface sediments. *Nature*, **454**(7207), 991–994. 19

Löytynoja, Ari, et Goldman, Nick. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**(5883), 1632–1635. 26

Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W., et Springer, M. S. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature*, **409**(6820), 610–614. 48

Margulis, Lynn. 1970. *Origin of Eukaryotic Cells:Evidence and Research Implications for a Theory of the Origin and Evolution of Microbial, Plant, and Animal Cells on the Precambrian Earth.* Yale University Press. 45

Marlétaz, Ferdinand, Martin, Elise, Perez, Yvan, Papillon, Daniel, Caubit, Xavier, Lowe, Christopher J, Freeman, Bob, Fasano, Laurent, Dossat, Carole, Wincker, Patrick, Weissenbach, Jean, et Parco, Yannick Le. 2006. Chaetognath phylogenomics: a protostome with deuterostome-like development. *Curr Biol*, **16**(15), R577–R578. 48

Martin, W., et Müller, M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature*, **392**(6671), 37–41. 45

Matte-Tailliez, Oriane, Brochier, Céline, Forterre, Patrick, et Philippe, Hervé. 2002. Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol*, **19**(5), 631–639. 46

Mau, B., et Newton, M.A. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, **6**, 122–131. 38

Metropolis, N, Rosenbluth, A W, m N Rosenbluth, Teller, A H, et Teller, E. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092. 37

Mojzsis, S. J., Arrhenius, G., McKeegan, K. D., Harrison, T. M., Nutman, A. P., et Friend, C. R. 1996. Evidence for life on Earth before 3,800 million years ago. *Nature*, **384**(6604), 55–59. 21

Moreira, D., Guyader, H. Le, et Philippe, H. 2000. The origin of red algae and the evolution of chloroplasts. *Nature*, **405**(6782), 69–72. 48

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W., et Springer, M. S. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, **294**(5550), 2348–2351. 48

Notredame, C., Higgins, D. G., et Heringa, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302**(1), 205–217. 26

Pagel, M. 1999. Inferring the historical patterns of biological evolution. *Nature*, **401**(6756), 877–884. 33

Pagel, Mark, Meade, Andrew, et Barker, Daniel. 2004. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*, **53**(5), 673–684. 33

Philippe, H., et Forterre, P. 1999. The rooting of the universal tree of life is not reliable. *J Mol Evol*, **49**(4), 509–523. 42

Philippe, Hervé, Snell, Elizabeth A, Bapteste, Eric, Lopez, Philippe, Holland, Peter W H, et Casane, Didier. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol*, **21**(9), 1740–1752. 48

Preston, C. M., Wu, K. Y., Molinski, T. F., et DeLong, E. F. 1996. A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov. *Proc Natl Acad Sci U S A*, **93**(13), 6241–6246. 18

Qiu, Y. L., Lee, J., Bernasconi-Quadroni, F., Soltis, D. E., Soltis, P. S., Zanis, M., Zimmer, E. A., Chen, Z., Savolainen, V., et Chase, M. W. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature*, **402**(6760), 404–407. 48

Rannala, B., et Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*, **43**(3), 304–311. 38, 39

Rannala, Bruce, et Yang, Ziheng. 2007. Inferring speciation times under an episodic molecular clock. *Syst Biol*, **56**(3), 453–466. 42

Ranwez, Vincent, Delsuc, Frédéric, Ranwez, Sylvie, Belkhir, Khalid, Tilak, Marie-Ka, et Douzery, Emmanuel Jp. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol Biol*, **7**, 241. 48

Rashby, Sky E, Sessions, Alex L, Summons, Roger E, et Newman, Dianne K. 2007. Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proc Natl Acad Sci U S A*, **104**(38), 15099–15104. 20

Rasmussen, Birger, Fletcher, Ian R, Brocks, Jochen J, et Kilburn, Matt R. 2008. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature*, **455**(7216), 1101–1104. 20

Robert, F., et Chaussidon, M. 2006. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature*, **443**(7114), 969–972. 21, 23

Rodríguez-Ezpeleta, Naiara, Brinkmann, Henner, Burey, Suzanne C, Roure, Béatrice, Burger, Gertraud, Löffelhardt, Wolfgang, Bohnert, Hans J, Philippe, Hervé, et Lang, B. Franz. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr Biol*, **15**(14), 1325–1330. 45

Rodríguez-Ezpeleta, Naiara, Brinkmann, Henner, Roure, Béatrice, Lartillot, Nicolas, Lang, B. Franz, et Philippe, Hervé. 2007a. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.*, **56**(3), 389–399. 49

Rodríguez-Ezpeleta, Naiara, Brinkmann, Henner, Burger, Gertraud, Roger, Andrew J, Gray, Michael W, Philippe, Hervé, et Lang, B. Franz. 2007b. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Curr Biol*, **17**(16), 1420–1425. 48

Rosing, Minik T. 1999. 13C-Depleted carbon microparticles in >3700-Ma seafloor sedimentary rocks from west greenland. *Science*, **283**(5402), 674–676. 21

Saitou, N., et Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**(4), 406–425. 35

Sanger, F., Nicklen, S., et Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**(12), 5463–5467. 40

Savolainen, V., Chase, M. W., Hoot, S. B., Morton, C. M., Soltis, D. E., Bayer, C., Fay, M. F., de Bruijn, A. Y., Sullivan, S., et Qiu, Y. L. 2000. Phylogenetics of flowering plants based on combined analysis of plastid atpB and rbcL gene sequences. *Syst Biol*, **49**(2), 306–362. 48

Savolainen, Vincent, et Chase, Mark W. 2003. A decade of progress in plant molecular phylogenetics. *Trends Genet*, **19**(12), 717–724. 48

Schimel, Joshua. 2004. Playing scales in the methane cycle: from microbial ecology to the globe. *Proc Natl Acad Sci U S A*, **101**(34), 12400–12401. 18

Schopf, J. William. 2006. Fossil evidence of Archaean life. *Philos Trans R Soc Lond B Biol Sci*, **361**(1470), 869–885. 4, 19

Scott, C., Lyons, T. W., Bekker, A., Shen, Y., Poulton, S. W., Chu, X., et Anbar, A. D. 2008. Tracing the stepwise oxygenation of the Proterozoic ocean. *Nature*, **452**(7186), 456–459. 23

Shen, Y., Buick, R., et Canfield, D. E. 2001. Isotopic evidence for microbial sulphate reduction in the early Archaean era. *Nature*, **410**(6824), 77–81. 22

Soria-Carrasco, Victor, et Castresana, Jose. 2008. Estimation of Phylogenetic Inconsistencies in the Three Domains of Life. *Mol Biol Evol*, Aug. 47

Studier, J. A., et Keppler, K. J. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol*, **5**(6), 729–731. 35

Summons, R. E., Jahnke, L. L., Hope, J. M., et Logan, G. A. 1999. 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature*, **400**(6744), 554–557. 20

Summons, Roger E, Bradley, Alexander S, Jahnke, Linda L, et Waldbauer, Jacob R. 2006. Steroids, triterpenoids and molecular oxygen. *Philos Trans R Soc Lond B Biol Sci*, **361**(1470), 951–968. 20

Tamura, K. 1992. The rate and pattern of nucleotide substitution in Drosophila mitochondrial DNA. *Mol Biol Evol*, **9**(5), 814–825. 33

Tateno, Y., Takezaki, N., et Nei, M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol Biol Evol*, **11**(2), 261–277. 33

Thompson, J. D., Higgins, D. G., et Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**(22), 4673–4680. 26

Tomitani, Akiko, Knoll, Andrew H, Cavanaugh, Colleen M, et Ohno, Terufumi. 2006. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc Natl Acad Sci U S A*, **103**(14), 5442–5447. 45

Ueno, Yuichiro, Yamada, Keita, Yoshida, Naohiro, Maruyama, Shigenori, et Isozaki, Yukio. 2006. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature*, **440**(7083), 516–519. 21

Ward, Peter, Labandeira, Conrad, Laurin, Michel, et Berner, Robert A. 2006. Confirmation of Romer's Gap as a low oxygen interval constraining the timing of initial arthropod and vertebrate terrestrialization. *Proc Natl Acad Sci U S A*, **103**(45), 16818–16822. 22

Whelan, S., et Goldman, N. 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. 33

Wildman, Derek E, Uddin, Monica, Opazo, Juan C, Liu, Guozhen, Lefort, Vincent, Guindon, Stephane, Gascuel, Olivier, Grossman, Lawrence I, Romero, Roberto, et Goodman, Morris. 2007. Genomics, biogeography, and the diversification of placental mammals. *Proc Natl Acad Sci U S A*, **104**(36), 14395–14400. 48

Woese, C. R., et Fox, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, **74**(11), 5088–5090. 5, 18, 40

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*, **39**(3), 306–314. 103

Yang, Z., et Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol*, **14**(7), 717–724. 38

Yang, Z., et Roberts, D. 1995. On the Use of Nucleic Acid Sequences to Infer Branchings in the Tree of Life. *Mol. Biol. Evol.*, **12**(3), 451–458. 42, 44

Yang, Ziheng. 2006. *Computational Molecular Evolution*. Oxford University Press. 33

Yap, Von Bing, et Speed, Terry. 2005. Rooting a phylogenetic tree with nonreversible substitution models. *BMC Evol Biol*, **5**(1), 2. 42, 44

Zablen, L. B., Kissil, M. S., Woese, C. R., et Buetow, D. E. 1975. Phylogenetic origin of the chloroplast and prokaryotic nature of its ribosomal RNA. *Proc Natl Acad Sci U S A*, **72**(6), 2418–2422. 6, 45

Zhaxybayeva, O., Lapierre, P., et Gogarten, J. P. 2005. Ancient gene duplications and the root(s) of the tree of life. *Protoplasma*, **277**(1), 53–64. 42

Zuckerkandl, E., et Pauling, L. 1965a. *Evolving Genes and Proteins*. Academic Press, New York. Chap. Evolutionary divergence and convergence in proteins, pages 97–116. 42

Zuckerkandl, E., et Pauling, L. 1965b. Molecules as documents of evolutionary history. *J Theor Biol*, **8**(2), 357–366. 26

Appendices

I have added two articles I contributed to during my license in Lyon and during the first year of my master in Uppsala, in Sweden.
The first article (section 11.1) attempts to better characterize the large amounts of gene duplications that occured at the base of the vertebrate clade. It was already known that an intense phase of gene duplication occured in the chordate lineage after the split from cephalochordates (amphioxus), but then it was unknown when this phase ended. By systematically building and analysing phylogenies for all genes from chondrichthyans (sharks, rays, and chimaeras) that were present in databases at the time, we estimated that this phase of intense duplication ended before the chondrichthyans separated from other vertebrates.

The second article (section 11.2) focuses on reconstructing ancestral gene contents in alpha-Proteobacteria, the bacterial family that gave birth to mitochondria. To this end, I used maximum parsimony, and simply considered the number of genes present in extant genomes, not their sequences. Despite its very crude nature, this procedure allowed us to estimate that the ancestor of all alpha-Proteobacteria was a free-living aerobe, and permitted to see genome reductions at the origin of parasites, and genome expansions at the origin of plant-interacting species.

11.1 Genome duplications and sharks

Phylogenetic Dating and Characterization of Gene Duplications in Vertebrates: The Cartilaginous Fish Reference

Marc Robinson-Rechavi,[1] *Bastien Boussau, and Vincent Laudet*

Laboratoire de Biologie Moléculaire de la Cellule, UMR CNRS5161, Ecole Normale Supérieure de Lyon, Lyon, France

Vertebrates originated in the lower Cambrian. Their diversification and morphological innovations have been attributed to large-scale gene or genome duplications at the origin of the group. These duplications are predicted to have occurred in two rounds, the "2R" hypothesis, or they may have occurred in one genome duplication plus many segmental duplications, although these hypotheses are disputed. Under such models, most genes that are duplicated in all vertebrates should have originated during the same period. Previous work has shown that indeed duplications started after the speciation between vertebrates and the closest invertebrate, amphioxus, but have not set a clear ending. Consideration of chordate phylogeny immediately shows the key position of cartilaginous vertebrates (Chondrichthyes) to answer this question. Did gene duplications occur as frequently during the 45 Myr between the cartilaginous/bony vertebrate split and the fish/tetrapode split as in the previous approximately 100 Myr? Although the time interval is relatively short, it is crucial to understanding the events at the origin of vertebrates. By a systematic appraisal of gene phylogenies, we show that significantly more duplications occurred before than after the cartilaginous/bony vertebrate split. Our results support rounds of gene or genome duplications during a limited period of early vertebrate evolution and allow a better characterization of these events.

Introduction

Vertebrates originated in the lower Cambrian (Shu et al. 2001), and their diversification and morphological innovations have been attributed to large-scale gene or genome duplications at the origin of the group (Ohno 1970; Holland et al. 1994). These duplications are predicted to have occurred in two rounds, the "2R" hypothesis, although it may have been one genome duplication plus many segmental duplications (Gu, Wang, and Gu 2002; McLysaght, Hokamp, and Wolfe 2002; Panopoulou et al. 2003). An interesting prediction of this hypothesis is that most genes that are duplicated in all vertebrates should have originated during the same period (for a discussion of predictions of the model, see Durand [2003]). Gene phylogenies consistent with this model are predicted to contain most duplications during a given speciation interval. The comparison of gene complexes, such as hox (Holland et al. 1994; Force, Amores, and Postlethwait 2002) or MHC (Abi-Rached et al. 2002), between species chosen for their key positions in the phylogeny of chordates, thus consistently date a large number of gene duplications after the divergence between the amphioxus and vertebrates (fig. 1). The choice of complexes of linked genes limits the insight these studies bring into the evolution of the whole genome, because each group of linked genes only samples one locus. Studies of the distribution and age of duplicated genes in the whole human genome sequence have established that gene duplications were indeed a massive phenomenon at the origin of vertebrates (Gu, Wang, and Gu 2002; McLysaght, Hokamp, and Wolfe 2002). However, because of their reliance on only one complete genome from a chordate and their reliance on the molecular clock, these studies cannot be very precise with respect to the dating and to the order of events, although efforts were done to add more species to the gene trees. In a pioneering comparison of phylogenies of unlinked genes, the tree topologies obtained were inconsistent with a simple scenario of two rounds of tetraploidization (Hughes 1999), but no dating of events was proposed. Phylogenies of gene families from various chordates show similar numbers of duplications before and after the lamprey/hagfish/gnathostome split, but results are not explained simply by two tetraploidizations (Escriva et al. 2002). All of these results are consistent with periods of intensive gene duplication, rather than genome duplication (Gu, Wang, and Gu 2002), although a recent phylogenetic study challenges even this scenario (Friedman and Hughes 2003).

Overall, there is support for a large number of gene duplications after the divergence between cephalochordates and vertebrates (Panopoulou et al. 2003), both before and after the lamprey/hagfish/gnathostome split (Escriva et al. 2002). This possibility leaves an important question mark on the ending time of the duplication events, which could represent a punctual event or could have occurred gradually over a period of 160 to 300 Myr. Consideration of chordate phylogeny (fig. 1) immediately shows the key position of chondrichthyans: if the massive gene duplications occurred almost exclusively before or after the chondrichthyan (cartilaginous vertebrates)/teleostome (bony vertebrates) split, this event supports "rounds" of duplications during a limited period of early vertebrate evolution. Otherwise, if gene duplications are evenly spread over the period between the cephalochordate/vertebrate split and the actinopterygian/sarcopterygian split, there is no evidence for these "rounds," but rather for a long period during which duplication was more frequent than in sarcopterygian evolution. Most studies do not include chondrichthyans, with the exception of two genes linked to the MHC, which were shown to be duplicated before the divergence of chondrichthyans and teleostomes (Abi-Rached et al. 2002).

Lack of chondrichthyan genome data has led us to use

[1] Present address: Joint Center for Structural Genomics, University of California, San Diego, La Jolla.

Key words: shark, ray, genome duplication, 2R hypothesis, phylogeny, Chondrichthyes.

E-mail: marc@sdsc.edu.

Mol. Biol. Evol. 21(3):580–586. 2004
DOI: 10.1093/molbev/msh046
Advance Access publication December 23, 2003
Molecular Biology and Evolution vol. 21 no. 3

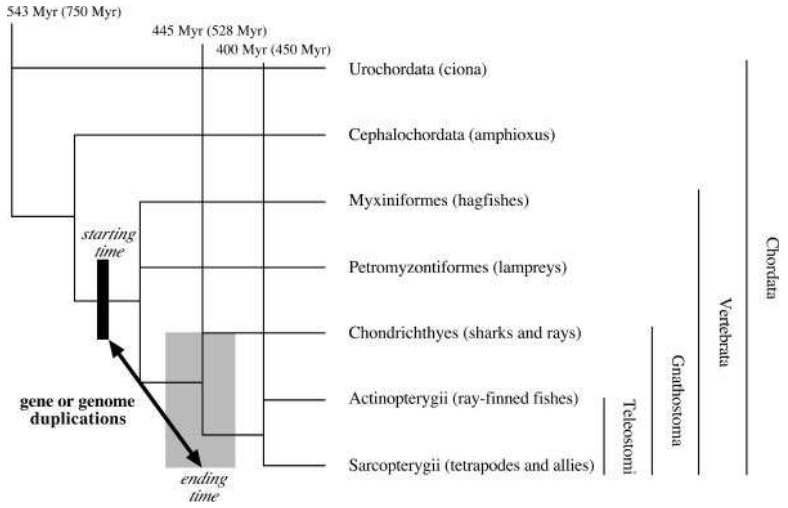


FIG. 1.—Possible timing of duplication events in chordate phylogeny. Schematic view of phylogenetic relations between chordates and possible timing of rounds of gene or genome duplication according to recent results (not including this work). The black bar represents relative confidence that duplications occurred essentially after the cephalochordate/vertebrate split, whereas the gray area represents the incertitude over the period when the duplication ended. Divergence dates are according to the fossil record (Samson, Smith, and Smith 1996; Shu et al. 1999; Zhu, Xiaobo, and Janvier 1999; Basden et al. 2000; Shu et al. 2001); molecular clock dates are shown in parentheses (Nikoh et al. 1997; Kumar and Hedges 1998). Although the topology (urochordates, [cephalochordates, vertebrates]) is well established, the corresponding dates of divergence are not known, apart from estimates of the date of apparition of chordates, given here as a conservative estimate of the first divergence among chordates.

the gene phylogeny approach to solve the question of when vertebrate-specific gene duplications did happen, by constructing phylogenetic trees of many protein-coding genes sequenced in Chondrichthyes. As mentioned above, if there were two major rounds of duplication, whether of genes or genomes, we would expect most gene families to show similar relative timing of speciation and duplication events. It should be noted that we are only interested in vertebrate-specific duplications here. Duplications that predate the chordate/arthropod/nematode split (approximately the origin of bilaterian animals), or more recent duplications such as frequently observed in actinopterygian fishes (Robinson-Rechavi et al. 2001), are outside the scope of this study.

Materials and Methods

Data Set

A first selection of gene families was done on Hovergen (Duret, Mouchiroud, and Gouy 1994) version 42 (April 2002), with the following criteria: at least one Chondrichthyes sequence, sequences from at least two Teleostome classes (to distinguish vertebrate specific and class specific duplications), and exclusion of mitochondrion-encoded genes. These criteria selected 149 gene families, as defined in Hovergen, including 415 chondrichthyan protein sequences. Protein alignments corresponding to the selected families were saved from Hovergen and checked using Seaview (Galtier, Gouy, and Gautier 1996). Outgroup sequences were added by Blast (Altschul et al. 1990) searches on Swissprot+TrEMBL (Boeckmann et al. 2003), excluding results from Vertebrata and from viruses, as implemented at PBIL (Perrière et al. 2003), and by Blast searches on the genome sequences of *Drosophila melanogaster* (Adams et al. 2000), *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998), *Ciona intestinalis* (Dehal et al. 2002), and *Anopheles gambiae* (Holt et al. 2002). Twelve gene families for which no outgroup sequence could be reliably identified were excluded.

Gene families with duplications predating the arthropod/nematode/chordate divergence (fig. 2*A*) were split into subfamilies, which were then evaluated separately for vertebrate-specific duplications. In cases of a vertebrate gene without any known mammalian ortholog, additional Blast searches were done on the human genome (International Human Genome Sequencing Consortium 2001). In all Blast searches, an expect value of 0.01 and the default filter for repeated sequences were used, and potential new genes were assessed for relevance to our study by a phylogenetic analysis. Once gene trees were built (see below), 86 gene families were found to yield phylogenies that could not be interpreted for dating of events at the origin of vertebrates (see Results). Notably, insufficient phylogenetic resolution was diagnosed when the gene tree was strongly inconsistent with the expected species phylogeny (for example, lamprey grouping with chicken and mammals not monophyletic [NPY gene family]) with very low bootstrap support (i.e., under 50%).

Phylogeny

All analyses were done using only complete sites (no gap, no X). When the inclusion of partial sequences led to less than 50 complete sites in the alignment, these sequences were excluded manually in Phylo_win (Galtier, Gouy, and Gautier 1996), taking care to keep representatives of each taxonomic group (i.e., actinopterygians, sarcopterygians, chondrichthyans, and outgroup) and of each paralog, as much as possible. Sequences that did not pass a χ^2 test for homogeneity of amino acid composition (as implemented in Tree-Puzzle [Schmidt et al. 2002])

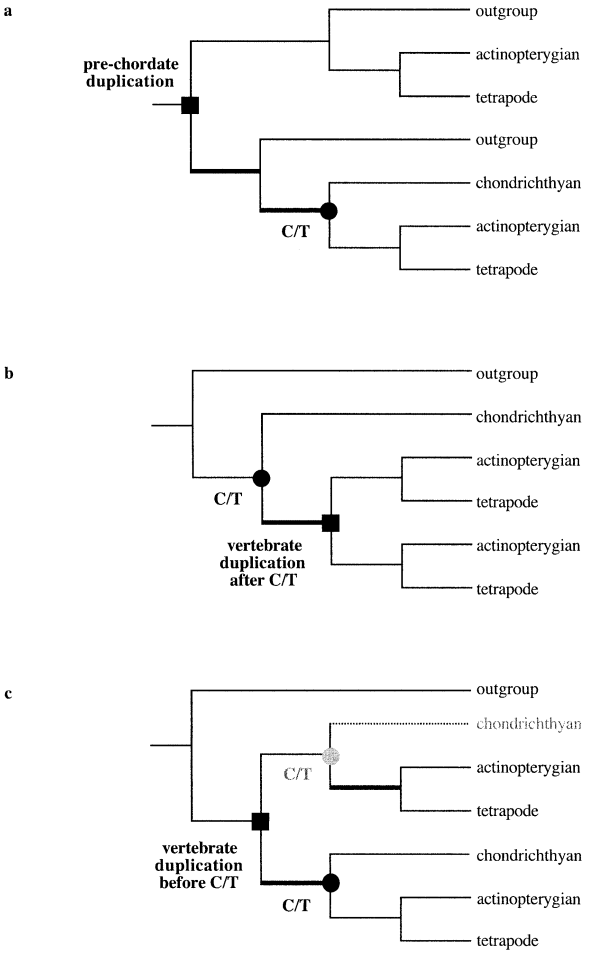


FIG. 2.—Classification of gene family phylogenies. Three schematic phylogenies, illustrating the possible interpretations of the order of the timing of gene duplications in a gene family. The taxon names represent gene sequences from these taxa, and "outgroup" represents sequences from nonvertebrate species.The branch(es) that should be tested for the classification of the gene family to be supported are in boldface. (*A*) No vertebrate specific duplication occurred, although gene duplications may (or may not) have occurred before the divergence of chordates from other animal lineages. (*B*) Vertebrate-specific gene duplication after the chondrichthyan/teleostome split. (*C*) Vertebrate-specific gene duplication before the chondrichthyan/teleostome split; the broken line indicates that the conclusion can be reached even if only one chondrichthyan homolog has been sequenced.

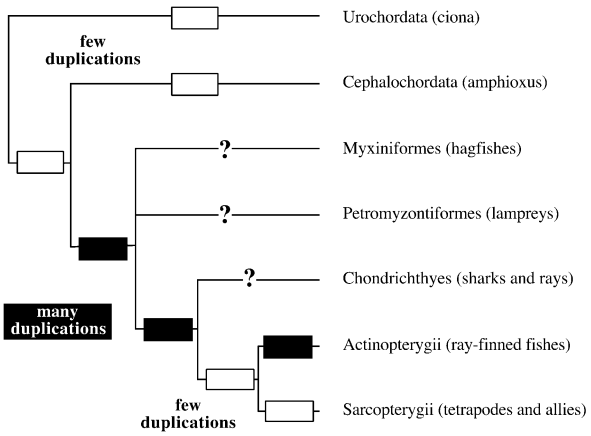


FIG. 3.—Gene duplication history in chordates. Present knowledge (including this work) of rounds of gene duplication mapped on the schematic view of phylogenetic relations between chordates. Black boxes represent characterized rounds of duplication, white boxes represent characterized periods with little accumulation of duplicate genes, and question marks represent lack of data to characterize duplications.

were excluded. This exclusion meant that six gene families no longer fulfilled the conditions set in terms of species sampling and were thus excluded from the data set. Trees were constructed using Neighbor-Joining (Saitou and Nei 1987) with distances corrected for multiple substitutions under a gamma model of rate heterogeneity (Yang 1996); the alpha parameter of the gamma model was estimated for each alignment by Tree-Puzzle version 5.1 (Schmidt et al. 2002) with eight rate categories, using default parameters. The following topologies were systematically compared by an SH likelihood test (Shimodaira and Hasegawa 1999), under the VT substitution model (Muller and Vingron 2000) with a γ model of rate heterogeneity, as implemented in Tree-Puzzle 5.1 (Schmidt et al. 2002): (1) species tree with no duplication (fig. 2*A*), (2) duplication after the chondrichthyan/teleostome split (fig. 2*B*), (3) duplication before the chondrichthyan/teleostome split (fig. 2*C*). When there were more than two vertebrate paralogs, all relative positions of the chondrichthyan/teleostome split and the duplications were compared; for example (Chondr (Teleos-α (Teleos-β, Teleos-γ))) versus (Teleos-α (Chondr (Teleos-β, Teleos-γ))) versus (Teleos-α (Teleos-β (Chondr, Teleos-γ))) versus ((Teleos-α, Chondr), (Teleos-β, Teleos-γ)) and so on. Results were considered supported if the likelihood of the favored topology was significantly higher than that of the best alternative topology (SH test; $P < 0.05$). Other results are classified as "not supported." It should be noted that we are only interested in the relative order of events of gene duplication and the chondrichthyan/teleostome split. Thus, teleost fish-specific duplications, as well as contradictions between gene phylogeny and teleostome phylogeny, as long as the latter were not statistically supported (they never were), were not taken into consideration to classify phylogenetic results, as far as they do not hamper interpretation of the trees. Moreover, when there were inaccuracies in teleostome phylogeny in the Neighbor-Joining tree, likelihood tests were performed under both the Neighbor-Joining and the species topology; significance of results was robust to the change.

Results

We selected gene families for the study in three steps: (1) selection on taxonomic criteria (sampling of cartilaginous and bony vertebrates, outgroup sequence); (2) manual consideration of phylogenetic trees, to assess whether the gene families are appropriate to the question being asked; and (3) evaluation of phylogenetic robustness. Notably, a total of 86 gene families were eliminated in step 2. The main causes limiting interpretation were (1) after splitting into vertebrate-specific subfamilies, some genes no longer fulfill the conditions set in terms of species sampling (typically the chondrichthyan sequence fell in a subtree with mammalian sequences and no other taxa); (2) very short sequences (NPY genes for example) with no

Table 1
Distribution of Duplication Histories of Gene Families

	Duplication Timing					
	None		Before C/T Split		After C/T Split	
Outgroup	Chordate	Other	Chordate	Other	Chordate	Total
Significant	0	0	16	3	0	19
Not significant	12	3	9	3	2	29
Total	15 (31%)		31 (65%)		2 (4%)	48

NOTE.—Numbers of gene families supporting each evolutionary history; no gene is counted twice. Phylogenetic support is noted as "significant" if the position of the chondrichthyan gene(s) is supported by a likelihood test ($P < 0.05$). The "C/T" split is the divergence between Chondrichthyes and Teleostomi. Chordate outgroups include ascidians, amphioxuses, lampreys, and hagfishes.

phylogenetic resolution; (3) extremely conserved sequences with no phylogenetic resolution (histones for example); (4) clustered multigene families for which conversion and recombination are well documented, typically from the immunological system; and (5) other genes with no phylogenetic resolution, such as hox genes, which include a very conserved homeodomain, with little information, the rest of the sequence being very divergent and with little information also (see a discussion in Force, Amores, and Postlethwait [2002]). It may be noted that while this selection mostly reduced the number of gene families used, splitting families with duplications predating the arthropod/nematode/chordate divergence increased the number of phylogenies analyzed (two additional "families" of proteasome beta subunit genes and one additional "family" of tyrosine phosphatase genes [see table 1 in Supplementary Material online]). Overall, the three steps of selection lead us from 149 gene families with cartilaginous and bony vertebrate homologous sequences to 48 gene families whose evolutionary history can be used to date duplication events at the origin of vertebrates (eliminated gene families in table 3 of Supplementary Material online), a figure very similar to the numbers of genes analyzed in recent studies using the same approach in other organisms (i.e., Langkjaer et al. 2003; Taylor et al. 2003).

Results for each gene family are detailed in the first table and the figures in the Supplementary Material online at www.mbe.oupjournals.org. Gene families with a duplication before the chondrichthyan/teleostome split (fig. 2*C*) clearly represent the majority of gene families we analyzed, including all 19 genes with significant phylogenetic resolution (table 1). Among the other 29 gene families, phylogenetic resolution is not significant at the chondrichthyan/teleostome divergence level (table 1). These include the only two gene families indicating a duplication after the chondrichthyan/teleostome split: a mannose-binding lectin, or tetranectin (HBG008208), and the PTP1D tyrosine phosphatase ("tyrosines phosphatases (1)" in the Supplementary Material online). Of note, a different result was found for PTP1D in a previous study that did not include all available mammalian sequences (Ono-Koyanagi et al. 2000). Finally, 15 genes show no evidence for any vertebrate-specific duplication. Our classification of these trees as "not supported" means that the species tree was not significantly more likely than other positions of chondrichthyans. This is consistent with a previous study in which individual nuclear genes had low power in solving the phylogenetic position of chondrichthyans (Martin 2001). The low phylogenetic resolution for the position of chondrichthyans among vertebrates is also consistent with the small divergence time between chondrichthyans and teleostomes reported in the fossil record (fig. 1). By contrast, the good phylogenetic resolution for the position of vertebrate-specific gene duplications may imply that the divergence time between these duplications and the chondrichthyan/teleostome split was important and that the duplications occurred early in vertebrate evolution.

It is possible that the observed distribution of gene duplications simply reflects the difference between the time intervals considered as "before the chondrichthyan/teleostome split" and "after the chondrichthyan/teleostome split." To test this, let us consider only the 27 gene families for which we have a vertebrate-specific duplication and a chordate outgroup (table 1: $16 + 9 + 2 = 27$), since they allow a more precise dating of events. If we use paleontological datings (fig. 1), the interval between chordate diversification and the chondrichthyan/teleostome split is 98 Myr, whereas the interval between this and the sarcopterygian/actinopterygian split is 45 Myr. Then we expect 31% (45/[98 + 45]) of vertebrate-specific gene duplications to be after the chondrichthyan/teleostome (C/T) split, under the assumption of a constant rate of gene duplication; the 95% confidence interval of this estimate is 14% to 49% ($f \pm 1.96\sqrt{var} = f(1-f)/N$; $f = 0.31$; $N = 27$). If we use molecular clock estimates of divergence dates (fig. 1), we expect 26% of gene duplications after C/T (confidence interval = 9.5% to 43%). The observed proportion of 7.4% (2/27) is significantly lower than expected by chance in either dating system (outside of the 95% confidence intervals). This conclusion holds true if we only use the 16 significantly supported phylogenies with a chordate outgroup (table 1): the observed proportion of duplications after the C/T split is 0%, whereas the expected value's confidence interval is either 8.7% to 54% (paleontological dates), or 4.5% to 47% (molecular clock dates). Thus, gene duplications are significantly less frequent after than before the chondrichthyan/teleostome split, taking into account evolutionary time.

Although our data set is not meant for detailed testing of duplication hypotheses in other branches of the tree, it is interesting to compare duplications that appear specific to either of the two major branches of teleostomes: out of 48 gene families, there are three with sarcopterygian-specific duplications and eight with actinopterygian-specific

duplications (see the second table in the Supplementary Material online at www.mbe.oupjournals.org), consistent with previous observations (Robinson-Rechavi et al. 2001). Interestingly, these more recent duplications concern 28% of the 32 gene families for which we have observed gene duplications ancestral to vertebrates but only 12.5% of the 16 gene families without vertebrate specific duplications.

Discussion

The "2R" hypothesis, modified from Ohno (1970), can be summarized by the idea that major duplication events occurred specifically in chordate genomes before the emergence of bony vertebrates. This hypothesis predicts that duplications should have occurred over a short period of time, in much greater numbers than in the previous or following periods. This prediction is shared by more recent hypotheses that there may have been one genome duplication and one major wave of segmental duplications (Gu, Wang, and Gu 2002; McLysaght, Hokamp, and Wolfe 2002; Panopoulou et al. 2003). The beginning time has been relatively well established, with studies showing that gene duplications occurred after the cephalochordate/vertebrate split and both before and after the gnathostome/jawless vertebrate split (Pennisi 2001; Wolfe 2001; Abi-Rached et al. 2002; Escriva et al. 2002; Gu, Wang, and Gu 2002; McLysaght, Hokamp, and Wolfe 2002; Panopoulou et al. 2003), but these studies did not set an ending time to these events. Given the prevalence of gene duplications in actinopterygian fishes (Wittbrodt, Meyer, and Schartl 1998; Robinson-Rechavi et al. 2001; Taylor et al. 2001), this raises the question of whether something specific really happened at the origin of vertebrates or whether gene duplications have been a common phenomenon throughout chordate evolutionary history, with the exception of sarcopterygians.

It is indeed noticeable that there has been no report of genome duplications ancestral to sarcopterygians (Pennisi 2001; Wolfe 2001; Durand 2003) or to any of the well-studied groups therein (e.g., tetrapodes, mammals, or sauropsids). Our own data are consistent with previous observations (Robinson-Rechavi et al. 2001; Taylor et al. 2001) that duplicate genes are significantly less abundant in sarcopterygians than in actinopterygians. Analysis of invertebrate chordate data also indicates that gene duplications are not abundant in these lineages (Dehal et al. 2002; Panopoulou et al. 2003).

Comparison of MHC-associated genes gave limited evidence for duplications before the chondrichthyan/teleostome split from two genes (Abi-Rached et al. 2002). Our results show that this pattern is general, with almost all vertebrate-specific gene duplications occurring before the chondrichthyan/teleostome split (table 1). This, added to all the previously published evidence, implies three waves of gene or genome duplications, two between the cephalochordate split and the chondrichthyan split and the other in actinopterygian fishes, separated by a period of "duplication calm" of about 45 Myr (which continued for 400 Myr in tetrapodes), which, although short, is significant. A major prediction of Ohno's (1970) original hypothesis, that of intense gene or genome duplication activity before the origin of vertebrates, is thus confirmed by the study.

Moreover, our results show that these gene duplications characterize all the jawed vertebrates and predict similar genetic complexity in sharks and rays as in tetrapodes. Consistent results are found for the evolution of hox clusters, which allow a direct connection between block duplications and morphological adaptations. Although hox genes are very poor phylogenetic markers, as illustrated by the difficulty in resolving the events that led to the different clusters of gnathostomes and lampreys (Force, Amores, and Postlethwait 2002; Irvine et al. 2002), partial sequences from the horn shark indicate that the duplications that led to four hox clusters in teleostomes occurred before the chondrichthyan/teleostome divergence (Kim et al. 2000). Moreover, horn shark and human hoxA clusters are remarkably conserved (Chiu et al. 2002). Thus, hox cluster analysis and our phylogenetic results are consistent in establishing no relation between gene duplications and the larger diversity of bony vertebrates than of cartilaginous vertebrates.

Although the basal branching of chondrichthyans among jawed vertebrates is considered extremely well supported by morphological and paleontological data (Janvier 1996), the analysis of complete mitochondrial sequences suggests a very different phylogeny, with chondrichthyans branching among bony ray-finned fishes (Actinopterygii) (Rasmussen and Arnason 1999). This surprising result has not been confirmed by any other source of data, and molecular phylogenies based on nuclear-encoded genes either are not informative (Martin 2001; this study) or strongly support the conventional branching position of chondrichthyans (Takezaki et al. 2003). In any case, our results show that vertebrate-specific gene duplications occurred before the divergence between chondrichthyans, actinopterygians, and sarcopterygians, whatever the order of these latter events.

Our results are at odds with a recent study that used a similar approach, dating gene duplications by their phylogenetic position relative to speciation events (Friedman and Hughes 2003). There are several differences between our methodology and that of Friedman and Hughes, but the main difference is the criterion for classifying gene duplications within speciation intervals. We consider genes to be duplicated within a given interval (i.e., between chordate diversification and the chondrichthyan/teleostome split) only if all relevant taxonomic groups (and thus speciations) are represented in the gene tree (i.e., a urochordate or a cephalochordate, a chondrichthyan, and a teleostome). Friedman and Hughes (2003) classify duplications as soon as they can be dated before or after one speciation. Moreover they used very distant dating points (i.e. the primate/rodent, amniote/amphibian, and deuterostome/protostome splits). It is unclear why they did not date duplications relative to the actinopterygian/sarcopterygian split, because this speciation would have been more relevant to the "2R" controversy, while taking advantage of genome data. As amphibians are the only lineage involved for which a genome sequence is not available, this may lead them

to include in the "before primate/rodent" category gene duplications that occurred before the amphibian/amniote split but for which they do not have amphibian sequences in the tree. This in turn may introduce a bias in their argument that the abundance of "before primate/rodent" versus "before amniote/amphibian" duplications is evidence against a peak of gene duplications at the origin of vertebrates. We believe that in our study, the division of the sequences into major taxonomic units, and our separation of the results according to the outgroup sequences used (table 1), preserve our results from such biases. Thus, differences in the conclusions between that study (Friedman and Hughes 2003) and ours probably reflect different sampling strategies.

An interesting side observation from our data set is that observations of gene duplications at the origin of vertebrates, and more recently in either the actinopterygian or sarcopterygian lineage, appear correlated. This may be the result of sampling; for example, better detection of duplications in more studied genes. Alternatively, it may indicate that the function of certain genes makes them more prone to persisting as duplicate copies. Such a tendency has indeed been recently shown in yeasts, where certain genes are retained independently as duplicates in different species (Hughes and Friedman 2003).

This study and other recent studies draw an increasingly precise picture of gene or genome duplication waves in chordates (fig. 3), although questions remain. Among the six branches of the chordate tree for which sufficient data are available, three are characterized by abundant preservation of duplicate genes, all of them in vertebrates. It has also been suggested on the basis of chromosome counts that polyploidy played an important part in lamprey evolution (Potter and Rothwell 1970). Of course it is probable that small-scale duplications have been continuous on all branches of the tree (Lynch and Conery 2000; Gu, Wang, and Gu 2002). However, large-scale duplications seem to have been frequent in vertebrate evolution, and the branches where they are absent, such as the origin of bony vertebrates, appear as the exception rather than the rule.

Acknowledgments

We thank Hector Escriva and Manolo Gouy for critical reading. This work was supported by the CNRS and the ENS Lyon.

Literature Cited

Abi-Rached, L., A. Gilles, T. Shiina, P. Pontarotti, and H. Inoko. 2002. Evidence of en bloc duplication in vertebrate genomes. Nat. Genet. **22**:22.

Adams, M. D., S. E. Celniker, R. A. Holt et al. (195 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185–2195.

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

Basden, A. M., G. C. Young, M. I. Coates, and A. Ritchie. 2000. The most primitive osteichthyan braincase? Nature **403**: 185–188.

Boeckmann, B., A. Bairoch, R. Apweiler et al. (12 co-authors). 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**:365–370.

Chiu, C.-H., C. Amemiya, K. Dewar, C.-B. Kim, F. H. Ruddle, and G. P. Wagner. 2002. Molecular evolution of the hoxA cluster in the three major gnathostome lineages. Proc. Natl. Acad. Sci. USA **99**:5492–5497.

Dehal, P., Y. Satou, R. K. Campbell et al. (87 co-authors). 2002. The draft genome of *Ciona intestinalis*: insights into Chordate and Vertebrate origins. Science **298**:2157–2167.

Durand, D. 2003. Vertebrate evolution: doubling and shuffling with a full deck. Trends Genet. **19**:2–5.

Duret, L., D. Mouchiroud, and M. Gouy. 1994. HOVERGEN: a database of homologous vertebrate genes. Nucleic Acids Res. **22**:2360–2365.

Escriva, H., L. Manzon, J. Youzon, and V. Laudet. 2002. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol. Biol. Evol. **19**:1440–1450.

Force, A., A. Amores, and J. H. Postlethwait. 2002. Hox cluster organization in the jawless vertebrate *Petromyzon marinus*. J. Exp. Zool. **294**:30–46.

Friedman, R., and A. L. Hughes. 2003. The temporal distribution of gene duplication events in a set of highly conserved human gene families. Mol. Biol. Evol. **20**:154–161.

Galtier, N., M. Gouy, and C. Gautier. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12**:543–548.

Gu, X., Y. Wang, and J. Gu. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat. Genet. **31**: 205–209.

Holland, P. W., J. Garcia-Fernandez, N. A. Williams, and A. Sidow. 1994. Gene duplications and the origins of vertebrate development. Development (suppl):125–133.

Holt, R. A., G. M. Subramanian, A. Halpern et al. (123 co-authors). 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science **298**:129–149.

Hughes, A. L. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. J. Mol. Evol. **48**:565–576.

Hughes, A. L., and R. Friedman. 2003. Parallel evolution by gene duplication in the genomes of two unicellular fungi. Genome Res. **13**:794–799.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Irvine, S. Q., J. L. Carr, W. J. Bailey, K. Kawasaki, N. Shimizu, C. T. Amemiya, and F. H. Ruddle. 2002. Genomic analysis of hox clusters in the sea lamprey *Petromyzon marinus*. J. Exp. Zool. **294**:47–62.

Janvier, P. 1996. Early vertebrates. Clarendon Press, Oxford.

Kim, C.-B., C. Amemiya, W. Bailey, K. Kawasaki, J. Mezey, W. Miller, S. Minoshima, N. Shimizu, G. Wagner, and F. Ruddle. 2000. Hox cluster genomics in the horn shark, *Heterodontus francisci*. Proc. Natl. Acad. Sci. USA **97**:1655–1660.

Kumar, S., and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature **392**:917–920.

Langkjaer, R. B., P. F. Cliften, M. Johnston, and J. Piskur. 2003. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature **421**:848–852.

Lynch, M., and J. S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151–1155.

Martin, A. 2001. The phylogenetic placement of chondrichthyes: inferences from analysis of multiple genes and implications for comparative studies. Genetica **111**:349–357.

McLysaght, A., K. Hokamp, and K. H. Wolfe. 2002. Extensive genomic duplication during early chordate evolution. Nat. Genet. **31**:200–204.

Muller, T., and M. Vingron. 2000. Modeling amino acid replacement. J. Comput. Biol. **7**:761–776.

Nikoh, N., N. Iwabe, K. Kuma et al. (11 co-authors). 1997. An estimate of divergence time of Parazoa and Eumetazoa and that of Cephalochordata and Vertebrata by aldolase and triose phosphate isomerase clocks. J. Mol. Evol. **45**:97–106.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, Heidelberg.

Ono-Koyanagi, K., H. Suga, K. Katoh, and T. Miyata. 2000. Protein tyrosine phosphatases from amphioxus, hagfish, and ray: divergence of tissue-specific isoform genes in the early evolution of vertebrates. J. Mol. Evol. **50**:302–311.

Panopoulou, G., S. Hennig, D. Groth, A. Krause, A. J. Poustka, R. Herwig, M. Vingron, and H. Lehrach. 2003. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res. **13**:1056–1066.

Pennisi, E. 2001. Genome duplications: the stuff of evolution? Science **294**:2458–2460.

Perrière, G., C. Combet, S. Penel et al. (11 co-authors). 2003. Integrated databanks access and sequence/structure analysis services at the PBIL. Nucleic Acids Res. **31**:3393–3399.

Potter, I. C., and B. Rothwell. 1970. The mitotic chromosomes of the lamprey, *Petromyzon marinus* L. Experientia **26**: 429–430.

Rasmussen, A. S., and U. Arnason. 1999. Molecular studies suggest that cartilaginous fishes have a terminal position in the piscine tree. Proc. Natl. Acad. Sci. USA **96**:2177–2182.

Robinson-Rechavi, M., O. Marchand, H. Escriva, P.-L. Bardet, D. Zelus, S. Hughes, and V. Laudet. 2001. Euteleost fish genomes are characterized by expansion of gene families. Genome Res. **11**:781–788.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Samson, I. J., M. M. Smith, and M. P. Smith. 1996. Scales of thelodont and shark-like fishes from the Ordovician of Colorado. Nature **379**:628–630.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502–504.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116.

Shu, D. G., L. Chen, J. Han, and X. L. Zhang. 2001. An early Cambrian tunicate from China. Nature **411**:472–473.

Shu, D. G., H. L. Luo, S. Conway Morris, X. L. Zhang, S. X. Hu, L. Chen, L. Han, M. Zhu, Y. Li, and L. Z. Chen. 1999. Lower Cambrian vertebrates from south China. Nature **402**:42–46.

Takezaki, N., F. Figueroa, Z. Zaleska-Rutczynska, and J. Klein. 2003. Molecular phylogeny of early vertebrates: monophyly of the agnathans revealed by sequences of 35 genes. Mol. Biol. Evol. **20**:287–292.

Taylor, J. S., I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer. 2003. Genome duplication: a trait shared by 22,000 species of ray-finned fish. Genome Res. **13**:382–390.

Taylor, J. S., Y. Van de Peer, I. Braasch, and A. Meyer. 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos. Trans. R. Soc. Lond. B Biol. Sci. **356**:1661–1679.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**:2012–2018.

Wittbrodt, J., A. Meyer, and M. Schartl. 1998. More genes in fish? Bioessays **20**:511–515.

Wolfe, K. H. 2001. Yesterday's polyploids and the mystery of diploidization. Nat. Rev. Genet. **2**:333–341.

Yang, Z. 1996. Among-site variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**:367–371.

Zhu, M., Y. Xiaobo, and P. Janvier. 1999. A primitive fossil fish sheds light on the origin of bony fishes. Nature **397**:607–610.

Hervé Phillippe, Associate Editor

Accepted October 25, 2003

11.2 Genome content evolution in the family of mitochondria

Computational inference of scenarios for α-proteobacterial genome evolution

Bastien Boussau, E. Olof Karlberg, A. Carolin Frank, Boris-Antoine Legault, and Siv G. E. Andersson†

Department of Molecular Evolution, Evolutionary Biology Center, Uppsala University, S-752 36 Uppsala, Sweden

Edited by Stanley Falkow, Stanford University, Stanford, CA, and approved May 18, 2004 (received for review February 11, 2004)

The α-proteobacteria, from which mitochondria are thought to have originated, display a 10-fold genome size variation and provide an excellent model system for studies of genome size evolution in bacteria. Here, we use computational approaches to infer ancestral gene sets and to quantify the flux of genes along the branches of the α-proteobacterial species tree. Our study reveals massive gene expansions at branches diversifying plant-associated bacteria and extreme losses at branches separating intracellular bacteria of animals and humans. Alterations in gene numbers have mostly affected functional categories associated with regulation, transport, and small-molecule metabolism, many of which are encoded by paralogous gene families located on auxiliary chromosomes. The results suggest that the α-proteobacterial ancestor contained 3,000–5,000 genes and was a free-living, aerobic, and motile bacterium with pili and surface proteins for host cell and environmental interactions. Approximately one third of the ancestral gene set has no homologs among the eukaryotes. More than 40% of the genes without eukaryotic counterparts encode proteins that are conserved among the α-proteobacteria but for which no function has yet been identified. These genes that never made it into the eukaryotes but are widely distributed in bacteria may represent bacterial drug targets and should be prime candidates for future functional characterization.

Fundamental questions subjected to much debate concern the extent to which microbial genomes are related by vertical descent versus horizontal gene transfer (1–5). A direct approach to address these questions is to estimate frequencies of deletions/duplications and horizontal gene transfers for closely related species and compare these estimates with estimates of nucleotide substitution rates. The α-proteobacteria provide an excellent model system for such studies because genome size variation in this subdivision spans the entire size range for bacteria, from 1 Mb in *Rickettsia* spp. to >9 Mb in *Bradyrhizobium japonicum* (6–12). Furthermore, there is an amazing variation in lifestyle characteristics in this subdivision, including both obligate (*Rickettsia* and *Wolbachia*) and facultative (*Bartonella* and *Brucella*) intracellular bacteria as well as soil-borne plant symbionts and pathogens (*Sinorhizobium*, *Agrobacterium*, and *Bradyrhizobium*), which enables correlations between gene contents and lifestyle features to be examined.

The α-proteobacterial group has also attracted much interest because one of its descending lineages is thought to be the ancestor of mitochondria (13, 14). The acquisition of mitochondria represents one of the earliest and most extreme cases of horizontal gene transfer events known in the history of life. Phylogenetic studies suggest that ≥630 eukaryotic genes were transferred from the α-proteobacteria to the eukaryotes, including many genes coding for modern mitochondrial protein functions (15). For the majority of mitochondrial proteins, however, no bacterial homologs were identified, indicating that they were derived from nuclear, eukaryotic genomes via intragenomic duplication and sequence divergence (14–16).

Based on results from pairwise genome comparisons, it has been suggested that there is a correlation between genome size alterations, microbial population sizes, and growth habitats (17). For example, it has been shown that free-living bacterial species of large population sizes accumulate insertion/deletion and rearrangement mutations relative to nucleotide substitutions at much higher frequencies than host-dependent bacteria of small population sizes, in which the influence of horizontal gene transfers has been negligible (17). Algorithms for mapping the presence and absence of genes onto inferred species trees in multiple genome comparisons (18, 19) have been used to reconstruct ancestral gene sets and to obtain estimates of the flow of genes along each of the individual branches. By using such approaches, >500 genes have been assigned to the last universal common ancestor (LUCA) (19), and 2,000 genes have been assigned to the ancestor of the Archaea (18).

In this study, we used the α-proteobacteria as a model system to examine the contents of ancestral genomes along with the evolutionary basis for genome size differences. Our results suggest that the α-proteobacterial ancestor contained several thousand genes and was metabolically highly versatile. The flux of genes along the individual branches of the tree highlights the role of the auxiliary chromosomes as mediators of genome size expansions and contractions in response to alterations in environmental conditions.

Materials and Methods

Genome Analysis. The sizes and GenBank accession numbers of α-proteobacterial genomes included in this analysis are given in Table 1. The assignment of functional categories for proteins in *Rickettsia prowazekii*, *Rickettsia conorii*, *Brucella melitensis*, *Brucella suis*, *Caulobacter crescentus*, *Agrobacterium tumefaciens*, *Sinorhizobium meliloti*, and *Mesorhizobium loti* was taken from the Institute for Genomic Research (www.tigr.org). Uncategorized proteins and proteins from *Bartonella henselae*, *Bartonella quintana*, and *B. japonicum* were assigned a functional category according to the best hit in similarity searches using BLASTP ($E < 1 \times 10^{-10}$) against all classified proteins from The Institute for Genomic Research (www.tigr.org). Additional proteobacterial genomes included as outgroups in the analyses were *Campylobacter jejuni* (NC_002163), *Escherichia coli* (NC_000913), *Helicobacter pylori* (NC_000913), *Pseudomonas aeruginosa* (NC_002516), *Ralstonia solanacearum* (NC_003296), *Salmonella typhimurium* (NC_003197 and NC_003277), and *Xylella fastidiosa* (NC_002490).

Phylogenetic Inference. The species phylogeny was estimated by using a data set of concatenated proteins that were selected on the basis that they are encoded by genes that are located in segments with largely conserved gene order structures in *B. henselae*, *B. quintana*, *B. melitensis*, *A. tumefaciens*, *S. meliloti*, and *M. loti* (see Fig. 6, which is published as supporting information on the PNAS web site). Homologs of the selected proteins *B. quintana* were inferred by BLASTP (20) searches ($E <$

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: LUCA, last universal common ancestor; COGs, Clusters of Orthologous Groups; BeT, symmetric best hit.

†To whom correspondence should be addressed at: Department of Molecular Evolution, Norbyvägen 18C, S-752 36 Uppsala, Sweden. E-mail: siv.andersson@ebc.uu.se.

www.pnas.org/cgi/doi/10.1073/pnas.0400975101

Table 1. α-Proteobacterial species included in the reconstruction analysis

Species	Total size, Mb	GenBank accession no. (size, Mb)
R. prowazekii	1.1	NC_000963 (1.1)
R. conorii	1.3	NC_003103 (1.3)
W. pipientis	1.3	NC_002987 (1.3)
B. quintana	1.6	BX897700 (1.6)
B. henselae	1.9	BX897699 (1.9)
B. melitensis	3.3	NC_003317 (2.1), NC_003318 (1.2)
B. suis	3.3	NC_004310 (2.1), NC_004311 (1.2)
C. crescentus	4.0	NC_002696 (4.0)
R. palustris	5.5	NC_005296 (5.5)
A. tumefaciens	5.6	NC_003062 (2.8), NC_003063 (2.1), NC_003064 (0.5), NC_003065 (0.2)
S. meliloti	6.7	NC_003047 (3.6), NC_003037 (1.4), NC_003078 (1.7)
M. loti	7.6	NC_002678 (7.0), NC_002679 (0.4), NC_002682 (0.2)
B. japonicum	9.1	NC_004463 (9.1)

1×10^{-20}) against the protein data set of each α-proteobacterial genome. To exclude paralogs we included in the analysis only genes without a second BLAST hit with an *E* value of $<1 \times 10^{-20}$. Another selection criteria for inclusion used was that orthologs should be present in at least 12 of the 20 taxa, resulting in a final set of 38 proteins (Table 3, which is published as supporting information on the PNAS web site).

The alignment was performed by using CLUSTALW (21) on individual protein sequences that were later concatenated. Maximum-likelihood phylogenies were constructed by using PHYML (version 2.1 beta) (22) assuming the Jones–Taylor–Thornton model of protein evolution and four γ-distributed rate categories with the α parameter and proportion of invariable sites estimated from the data. To assess the variation in the data, 100 bootstrap replicates were generated from the data set with SEQBOOT from the PHYLIP 3.5c package (J. Felsenstein, Department of Genetics, University of Washington, Seattle). Maximum-likelihood trees were estimated from the bootstrap matrices as described above, and a majority-rule consensus tree was generated from them by using CONSENSE, also from the PHYLIP 3.5C package.

Inference of Ancestral Gene Sets. The homologous groups were created by using the Clusters of Orthologous Groups (COGs) database (23) in its 66-genomes version. Proteomes classified in COGs were retrieved from the COGs database. Six unclassified proteomes (*B. henselae*, *B. quintana*, *B. suis*, *B. japonicum*, *Rhodopseudomonas palustris*, and *Wolbachia pipientis*) were assigned COGs according to the following procedure: the proteins in each unclassified proteome were used as first queries and then databases in separate BLAST searches with all proteomes in the COGs database. The unclassified proteins were added to the COG to which it had the highest number of symmetric best hits (BeTs) and BeTs >1. Because this procedure expanded the COGs, the same was done for all the unclassified proteins from the other species so as to also include proteins with BeTs to the newly assigned proteins. New clusters were then created from uncategorized proteins forming triangles of BeTs as described in ref. 23. Finally, clusters containing only two proteins were made from linear BeT relations, after which the remaining proteins were included as single genes.

The most parsimonious scenarios of α-proteobacterial genome evolution and the α-proteobacterial ancestor were reconstructed by character mapping by using generalized parsimony as implemented in PAUP* (version 4.0b10 for Unix) (24) on a rooted species tree, with ACCTRAN (accelerated transformation) (see Fig. 3) and DELTRAN (delayed transformation) (Fig. 7, which is published as supporting information on the PNAS web site) options for parsimony analysis. Fig. 3 shows the results for penalties for duplications, deletions, and gene genesis of 1, 1, and 5, respectively. The selection of penalty values and results obtained for different penalty values are described in Fig. 7.

The ancestral proteomes were inferred separately for protein families assigned to auxiliary (mega-COG) and main (main-COG) chromosomes. The criteria for inclusion in the mega-COG family were that ≥30% of the protein members were encoded on auxiliary replichores or symbiosis islands in the Rhizobiales. By using these criteria, 43% of the proteins encoded by the auxiliary replichores and 6% of chromosomally encoded proteins were members of the mega-COG families on average. Because many of the species-specific genes are located on the auxiliary replichores, we used the complete α-proteobacterial proteome for this analysis. The gene content of the inferred α-proteobacterial ancestral genome was compared with the estimated gene content of protomitochondria (15) and the LUCA (19) by using the presence or absence of a COG rather than the absolute numbers of genes.

Results and Discussion

Gene Function of α-Proteobacterial Genomes. To explore expansions in gene function with genome size for the α-proteobacteria (Table 1), we examined gene content statistics for 14 functional categories (Fig. 1). The relationships between gene content and genome size can be approximated with linear functions, with slopes ranging from four genes per megabase for basic information processes such as transcription and translation to >80 genes per megabase for energy metabolism, transport, and regulatory functions. Functional categories associated with environmental interactions (e.g., transport and regulation) were found to be the most variable among bacteria with different lifestyles. For example, the small genomes of obligate and facultative intracellular parasites have only a few regulatory and transport genes, whereas the larger genomes of free-living soil bacteria that alternate between environments of different nutritional quality contain hundreds of such genes. A rapid increase in the number of regulatory genes in relation to gene content has been observed (25, 26) and may be a general feature of all bacterial genomes.

Extrapolation to the intercept of the *y* axis provides a measure of the minimal set of genes shared among the α-proteobacteria, which here is estimated to 250 genes (Table 4, which is published as supporting information on the PNAS web site). This set includes ≈200 genes for DNA, RNA, and protein biosynthesis and another 40 genes for nucleotide and cofactor biosynthesis. This is comparable with the minimal set of core genes in endosymbiotic bacteria (27) as well as to minimal gene numbers inferred by computational approaches (28) and experimental knockout mutants of *Bacillus subtilis* (29).

The Species Tree for α-Proteobacteria. To place the dramatic shifts in genome size in an evolutionary context, we needed an underlying reliable species tree onto which the gene sets could be mapped. Because a few of the divergence nodes were not conclusively resolved in our rRNA tree (data not shown), we inferred the tree topology by using concatenated protein sequences (Fig. 2). To minimize topology inconsistencies caused by horizontal gene transfer and gene paralogy, we selected for this analysis a set of 38 genes sampled from regions with conserved gene order structures in the Rhizobiales (Fig. 6 and Table 3).

The phylogenetic tree (Fig. 2), constructed by using the maximum-likelihood method, provided strong support for a clustering of the Rhizobiales to the exclusion of the more early diverging lineages *B. japonicum*, *C. crescentus*, and the Rickettsiales. The two *Bartonella* species formed a clade with *Brucella*

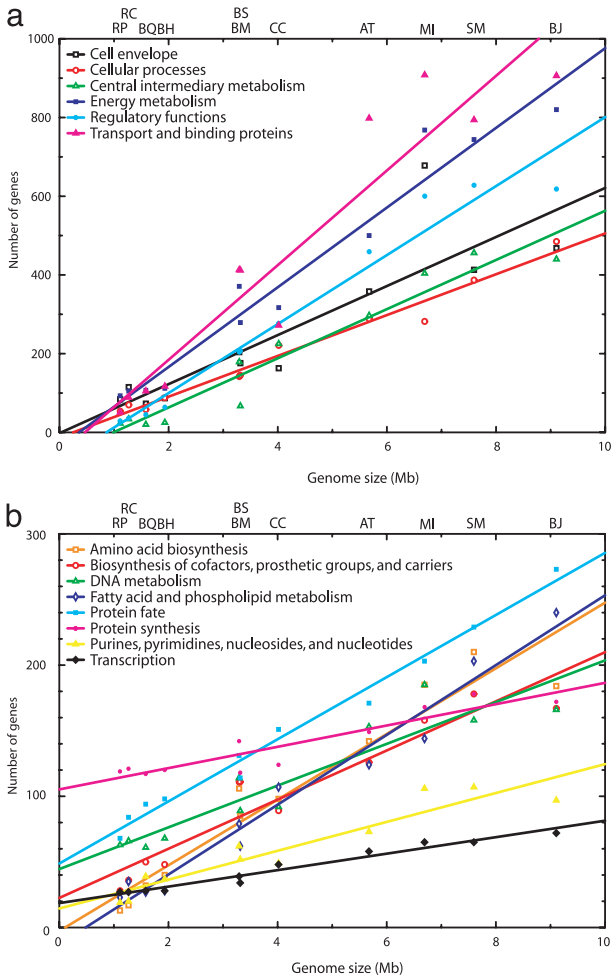


Fig. 1. Plot of genome size against gene content for each of the functional categories. RP, *R. prowazekii*; RC, *R. conorii*; BQ, *B. quintana*; BH, *B. henselae*; BM, *B. melitensis*; BS, *B. suis*; CC, *C. crescentus*; AT, *A. tumefaciens*; SM, *S. meliloti*; ML, *M. loti*; and BJ, *B. japonicum*. See Table 1 for genome sizes. The data were separated into two sections (*a* and *b*) to prevent overcrowding.

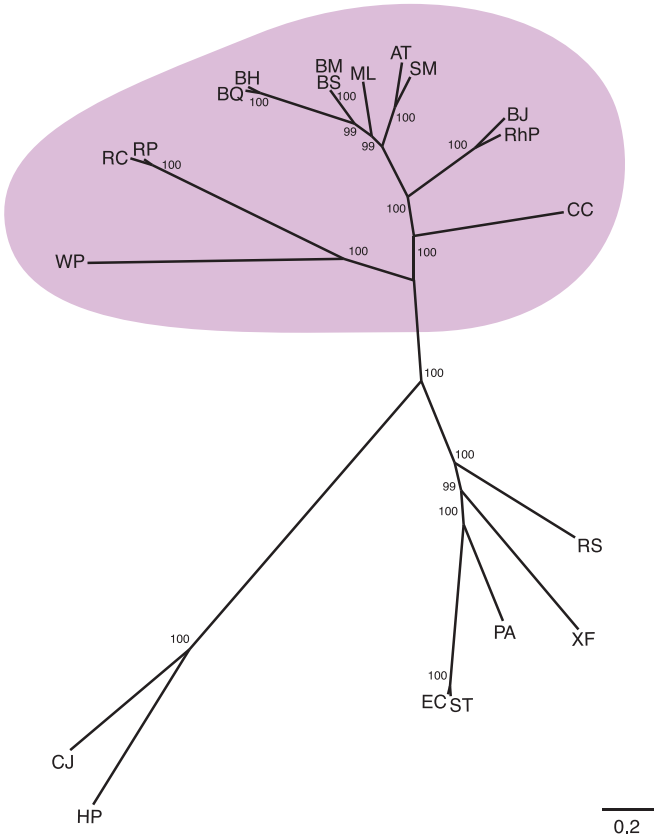


Fig. 2. Phylogenetic relationship of 13 α-proteobacterial species (highlighted by the purple background) with 7 species from other proteobacterial subdivisions as outgroups. The topology, branch lengths, and bootstrap support are according to maximum-likelihood reconstructions with the Jones–Taylor–Thornton + 4ΓI model. Similar results were obtained with the neighbor-joining method and after removal of positions with gaps. A list of genes used for the phylogenetic reconstructions is given in Table 5. Abbreviations for species names are as described in the legend to Fig. 1 with the addition of the following taxa: WP, *W. pipientis*; RhP, *R. palustris*; CJ, *C. jejuni*; EC, *E. coli*; HP, *H. pylori*; PA, *P. aeruginosa*; RS, *R. solanacearum*; ST, *S. typhi*; and XF, *X. fastidiosa*.

with high bootstrap support, as did also *A. tumefaciens* and *S. meliloti*, which formed a separate clade. The position of *M. loti* was placed with high support (>90%) close to the root of the *Bartonella/Brucella* clade. However, the branches separating *M. loti* from its neighboring clades are very short and the placement of *M. loti* in the tree was found to be sensitive both to the methods used and to the genes and species sampled (data not shown). For all other divergences, the tree topology was robust. The branching order depicted in Fig. 2 represents our best estimate of the underlying species tree.

Computational Inference of Ancestral Gene Sets. We inferred ancestral α-proteobacterial proteomes and estimated the number of gene losses, duplications, and genesis events along each branch of the topology shown in Fig. 2 with character mapping using generalized parsimony (Figs. 3 and 7). Following the routines of previous work (18, 19), we included in the analysis proteins already classified in the COGs database (23) along with proteins encoded by genomes not yet incorporated in the COGs database but related to existing COGs by BeTs. This process resulted in a first data set of 56,337 proteins, to which we added 384 COGs containing proteins not related to any existing COGs but present in three or more species and internally related by BeTs. With the inclusion of these proteins, the data set amounted to 58,171 proteins, and the α-proteobacterial ancestral proteome was estimated to 3,300 proteins (Fig. 3*a*). The remaining proteins were assigned into single or linear protein COGs, which resulted in a data set that included all 73,658 proteins and yielded an ancestral proteome of >5,000 proteins (Fig. 3*b*). Because some of the species-specific genes may be rapidly evolving or incorrectly annotated as genes, their inclusion probably results in an overestimate of the ancestral proteome size (Fig. 3*b*), just as their exclusion may yield an underestimate (Fig. 3*a*). Thus, we define the lower and upper boundaries of the ancestral α-proteobacterial proteome to 3,000 and 5,000 proteins, respectively.

Metabolic Expansions and Contractions. The analyses of gene content alterations at the branches of the tree revealed two major trends that are observed irrespectively of the different data sets and methods used (Fig. 4). First, massive genome size expansions accompanied the divergence of the plant-associated Rhizobiales, particularly the evolution of *M. loti* and *B. japonicum*. There seems to have been a gradual increase of genes encoding transcriptional regulators and proteins involved in the transport and metabolism of amino acids, nucleotides, carbohydrates, coenzymes, lipids, inorganic ions, and secondary metabolites. These expansions argue in favor of ancestral cells being visited

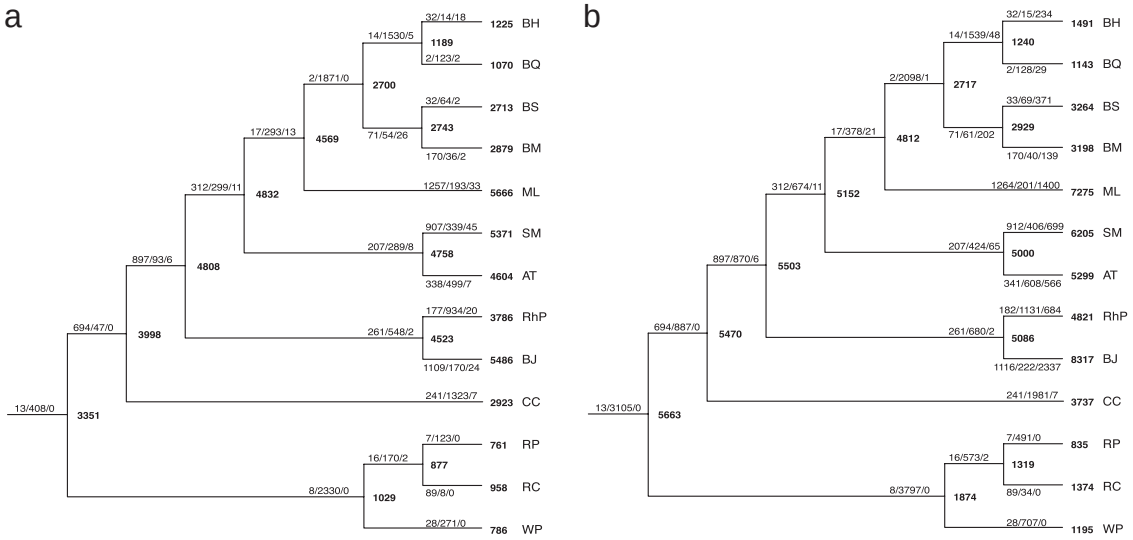


Fig. 3. Inference of deletions/duplications and gene-genesis events based on the α-proteobacterial tree was made by using different clustering levels and penalty values. The inference was based on proteins already classified in COGs (23) to which we added COGs containing proteins in three or more species internally related by best hits (58,171 proteins in total) (*a*) and the complete set of proteins (73,658 proteins in total) (*b*). Inference of gene contents was made by using the ACCTRAN option for parsimony analysis in PAUP* with penalties for duplication, deletion, and gene genesis set to 1, 1, and 5, respectively. Numbers along branches refer to the number of duplications/losses/genesis, respectively. Numbers at nodes refer to the putative number of genes in the inferred genome at the node. Outgroup sequences are as described for Fig. 2, but they were pruned from the tree shown here. Abbreviations for species names are as described in the legends to Figs. 1 and 2.

by highly dynamic plasmids that introduced novel genes by duplication and/or genesis, some of which were maintained selectively in response to the increased use of soil compounds and the refined interactions with the progenitors of modern plant cells.

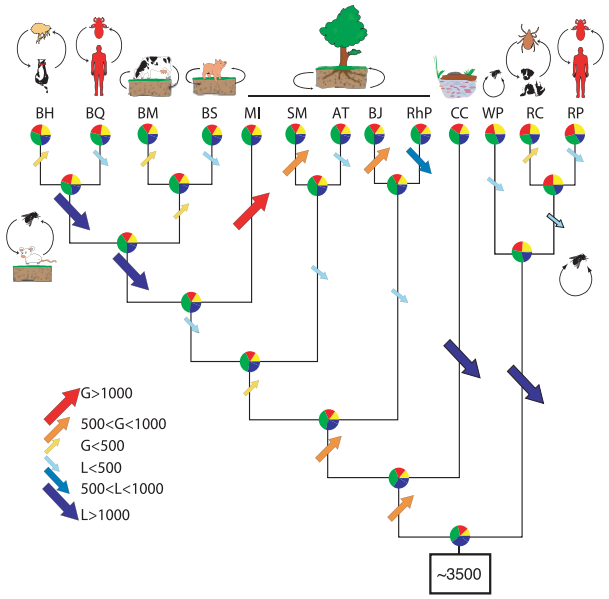


Fig. 4. Net gene loss or gain throughout the evolution of the α-proteobacterial species. Arrows pointing upward indicate net gains of genes (G), and arrows pointing downward indicate net losses of genes (L). Colors and sizes of arrows refer to the net number of genes gained or lost at each branch. Colors of circles refer to the relative fraction of genes assigned to the different functional groups in the modern and inferred genome at the node. Yellow, information storage and processing; green, metabolism; red, cellular processes; blue, poorly characterized. Clustering groups and estimated frequencies are as described for Fig. 3*a*. Abbreviations for species names are as described in the legends to Figs. 1 and 2.

Extreme reductions of size occurred twice independently: in the ancestor of the obligate intracellular lineages *Rickettsia* and *Wolbachia* and in the ancestor of the facultative intracellular lineages *Bartonella* and *Brucella*. These losses have largely affected protein families for transcription regulation, transport, and metabolism of amino acids, nucleotides, carbohydrates, lipids, and other small molecules. Particularly notable is the independent loss of genes involved in secretory pathways, pilus assembly, and flagellar biosynthesis. The loss of genes associated with the transition from interactions with plants to animals in the ancestor of *Bartonella* and *Brucella* was not balanced by a corresponding gain of genes; no genes have homologs solely in *Bartonella* and *Brucella* ($E < 0.001$).

The number of genes eliminated before the split of *Rickettsia* and *Wolbachia* was estimated to 2,300–3,800 genes, as compared with ≈200–700 lost genes per lineage after the split (Fig. 3). The inverse correlation between gene loss and branch lengths for this part of the tree (compare Figs. 2 and 3) makes the lower frequency of gene-elimination events in recent times all the more striking. On average, the ratio of deletions to nucleotide substitutions was 25-fold higher before the split of *Rickettsia* and *Wolbachia*. A high frequency of gene loss relative to nucleotide substitutions was also observed immediately before the emergence of the intracellular lineages *Bartonella* and *Brucella*, which is reminiscent of the more rapid loss of genes at an early stage of genome reduction in aphid endosymbiont lineages, followed by genomic stasis (17). Overall, we observed no correlation between frequencies of amino acid substitutions and gene loss ($r^2 = 0.14$), gene duplication ($r^2 = 0.02$), or gene genesis ($r^2 = 0.05$), indicating dramatically different fixation rates for these mutations in the different lineages over time.

Gene Flux on Chromosomes and Auxiliary Replicons. Many species in the Rhizobiales contain auxiliary chromosomes (Table 1) that are characterized by less gene synteny than the main chromosomes (Fig. 6). To quantify the differences in mutational rates and patterns for genes located on different replicons, we inferred ancestral proteomes separately for COGs assigned to the aux-

EVOLUTION

iliary replicons (mega-COG) versus those assigned to the main chromosomes (main-COG). We classified a COG as a mega-COG if >30% of its protein members were encoded on an auxiliary replicon in *A. tumefaciens*, *Brucella* spp., *S. meliloti*, or on the symbiosis islands in *M. loti* and *B. japonicum*. In total, we classified 13% of the COGs as mega-COGs, which corresponds to 2,349 COGs (8,662 proteins) out of the complete set of 17,669 COGs (73,658 proteins) included in the analysis.

The results showed that 20–24% of the losses that occurred immediately before the *Bartonella*/*Brucella* divergence was associated with mega-COGs (Fig. 8, which is published as supporting information on the PNAS web site). Likewise, a substantial fraction of the identified duplications involved proteins in mega-COG families, as observed for example on the branch leading to the Rhizobiales (23%) and also on the branch separating these from *R. palustris* and *B. japonicum* (55%). In the terminal branches for *S. meliloti* and *A. tumefaciens*, all three types of mutational events were frequent for proteins classified in the mega-COG family, including 30% of duplications, 25% of losses, and 60% of gene-genesis events. Overall, mega-COGs accounted for 21% of changes below the α-proteobacterial ancestor. Considering that the mega-COGs only account for 13% of all COGs, the relative frequencies of deletions, duplications, and gene genesis was considerably higher for proteins classified in these families. We speculate that the auxiliary replicons were derived from plasmids that expanded by reiterative processes of duplication/deletion and horizontal gene-transfer events in the Rhizobiales.

Inferred Metabolism of the α-Proteobacterial Ancestor. Our pathway analysis of the core ancestral gene set identified in all the analyses (Table 5, which is published as supporting information on the PNAS web site) suggests that it contained genes for glycolysis and a complete system for aerobic respiration, as expected for a unicellular organism that was well adapted to the aerobic environment. Notable was its broad biosynthetic capability and the presence of multiple genes for regulatory and transport functions. The analysis further identified genes for flagellar biosynthesis and type III and type IV secretion systems. Thus, the ancestor was probably a free-living, aerobic, and motile bacterium that had evolved elaborate communication mechanisms with other cells. Also present in the ancestor were genes for phage-related functions; however, these genes may incorrectly have been assigned to the ancestor because of multiple independent acquisitions of phage genes by horizontal gene transfer in some of the derived lineages.

A comparison of the α-proteobacterial ancestral genome with the gene content of the LUCA identified a small set of genes inferred to be present in the LUCA (13) but absent from our ancestral set. The number and identity of such genes depend on penalty values, but even for the highest penalty values it was observed that a set of genes, including those for homoserine kinase, uridine kinase, endonuclease IV, and glutamyl-tRNA reductase, were predicted to be present in the LUCA but were absent from the α-proteobacterial ancestor. These might have been lost before the divergence of the α-proteobacterial ancestor or, alternatively, been incorrectly assigned to the LUCA.

Comparing the α-Proteobacterial Ancestor with the Mitochondrial Ancestor. The endosymbiotic theory postulates that mitochondria evolved by massive gene loss and transfer of genes from the common ancestor to the nuclear genome of the host cell. A total of 630 orthologous groups display a close phylogenetic relationship between eukaryotes and α-proteobacteria (15). These represent a minimal estimate of the protomitochondrial proteome, because some gene transfers may have been missed because of weak phylogenetic signals and others may have been lost from the eukaryotic genomes included in the analysis. We compared

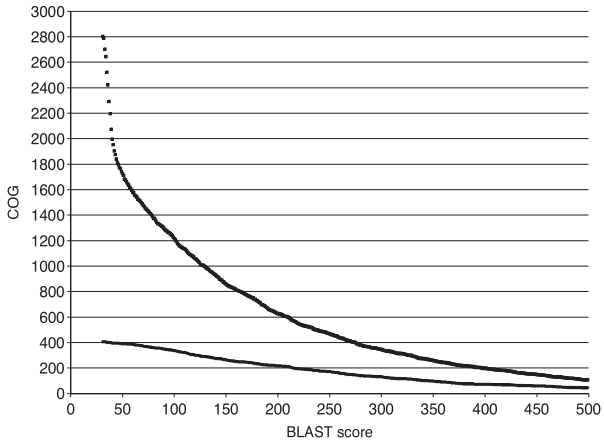


Fig. 5. Number of COGs in the α-proteobacterial ancestor (Fig. 3*a*) with sequence similarity to eukaryotic genes for different BLAST score values. Estimated number of COGs that shows similarity to eukaryotic genes in the inferred proteomes of the α-proteobacterial ancestor (upper curve) and the minimal protomitochondrial ancestor (lower curve) (15).

the 630 α-proteobacterial gene groups with the set of COGs inferred to be putatively present in the α-proteobacterial ancestor. The protomitochondrial set includes 487 genes in 412 COG-associated groups (15), all of which belong to the 3,300 genes in the >3,100 COGs of our ancestor (Fig. 3*a*). Of the 143 protomitochondrial groups not associated with a COG, 92 are represented in the ancestral gene pool. Most of the 51 groups missing from our data set consists of hypothetical proteins or proteins with unknown functions.

Phylogenetic analyses of rRNA sequences, protein subunits of the respiratory chain complexes, and concatenated protein alignment suggest that mitochondria evolved from the α-proteobacteria, with no evidence for multiple independent acquisitions (12, 13, 30–32). Although several studies have placed mitochondria as a deeply diverging sister clade near to the Rickettsiales (30–32), the exact position is still debated. Here, we consider the gene set of the reconstructed α-proteobacterial ancestor as an upper limit of the protomitochondrial proteome. To estimate how many of these ancestral genes may, at the most, have been transferred to the host nuclear genome, we selected the complete set of COGs present in the α-proteobacterial ancestor and used them as queries in sequence-similarity searches against eukaryotic genomes. As expected, the number of COGs showing significant sequence similarity to eukaryotic genes decreased with increasing BLAST scores from ≈1,700 (score ≥50) to 850 (score ≥150) (Fig. 5). The remaining 1,144 ancestral COGs without eukaryotic homologs (score ≤40) represent putative gene losses. The genes in these COGs display a broad taxonomic distribution in bacteria (data not shown), and surprisingly many (>45%) encode proteins of unknown or poorly characterized function (Table 2). Future functional analyses of these genes may provide the answers as to why these genes were not transferred to the eukaryotes.

Concluding Remarks

This study represents an attempt to quantify the different mutational changes that underlie genome size alterations in the α-proteobacteria. We observed no correlation between nucleotide substitution rates and fixation rates for mutations that affect genome contents. On the contrary, our results strongly suggest that the inferred frequencies of deletions, duplications, and horizontal gene transfers depend on population sizes and bacterial lifestyle features. In particular, the data support the

Table 2. Relative fraction of COGs in the α-proteobacterial ancestor (Fig. 3*b*) sorted according to broad functional categories

Functional category	+Hom*	Min†	−Hom*
Cellular processes	17	15	12
Information processes	15	15	6
Metabolism	45	53	14
Poorly characterized	20	17	45
New clusters‡	3	0	23

*Values are percentages of COGs in the α-proteobacterial ancestor with homologs (score $\geq$50) (+Hom) and without homologs (score $\leq$40) (−Hom) in eukaryotic genomes.

†Values are percentages of COGs in the minimal (Min) protomitochondrial genome (15) with homologs in eukaryotic genomes (score $\geq$50).

‡Uncategorized clusters created in this analysis.

suggested correlation between transitions to intracellular growth habitats and genome size reductions, with the highest frequencies of gene loss at early stages of the transition (17).

The stability of the main chromosomes of the Rhizobiales, displayed as segments with conserved gene synteny, contrasts with otherwise high substation rates and extensive gene-content differences. Expansions and contractions in the genomic repertoire have mostly affected genes involved in environmental interactions; these typically are located on the auxiliary replichores and evolve by very high turnover rates. It is possible that we have underestimated these rates at the internal branches of the tree because of multiple insertion/deletion events. High intrinsic rates for duplications/deletions and horizontal gene transfers may serve as an efficient mutational engine that enables rapid responses to alterations in the environmental conditions when subjected to strong selective pressures.

Although the estimated frequencies of duplication and gene-genesis events depend on the penalties assigned to these events, our study clearly demonstrates the importance of gene duplications for expanding and diversifying the metabolic and regulatory capacities of the bacterial cell. A consequence of high duplication and deletion rates is that the number of paralogous proteins may be much larger than previously anticipated. In effect, the many different protein variants do not necessarily trace back to one ancestral giant gene pool but may have arisen throughout evolution via reiterative processes of duplication and loss. The continuous generation of novel paralogs may provide one explanation for the difficulty to obtain congruent single gene trees in phylogenomic surveys (1–5).

Computational inference of ancestral genomes with refined models that account for the relative frequencies of the different types of mutational events in the different lineages will provide more detailed scenarios of genome size evolution in the α-proteobacteria and other bacterial subdivisions.

This research was supported by grants from the Swedish Research Council, the Swedish Foundation for Strategic Research, and the Wallenberg Foundation.

1. Doolittle, W. F. (1999) *Science* **284,** 2124–2128.
2. Snel, B., Bork, P. & Huynen, M. (1999) *Nat. Genet.* **21,** 108–110.
3. Sicheritz-Ponten, T. & Andersson, S. G. E. (2001) *Nucleic Acids Res.* **29,** 545–552.
4. Kurland, C. G., Canback, B. & Berg, O. G. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9658–9662.
5. Daubin, V., Moran, N. A. & Ochman, H. (2003) *Science* **301,** 829–832.
6. Andersson, S. G. E., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C. M., Podowski, R. M., Näslund, K., Eriksson, A.-S., Winkler, H. H. & Kurland, C. G. (1998) *Nature* **396,** 133–140.
7. Ogata, H., Audic, S., Renesto-Audiffren, P., Fournier, P. E., Barbe, V., Samson, D., Roux, V., Cossart, P., Weissenbach, J., Claverie, J. M. & Raoult, D. (2001) *Science* **293,** 2093–2098.
8. Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B. S., Cao, Y., Askenazi, M., Halling, C., *et al.* (2001) *Science* **294,** 2323–2328.
9. Wood, D. W., Setubal, J. C., Kaul, R., Monks, D. E., Kitajima, J. P., Okura, V. K., Zhou, Y., Chen, L., Wood, G. E., Almeida, N. F., Jr., *et al.* (2001) *Science* **294,** 2317–2322.
10. Galibert, F., Finan, T. M., Long, S. R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M. J., Becker, A., Boistard, P., *et al.* (2001) *Science* **293,** 668–672.
11. Kaneko, T., Nakamura, Y., Sato, S., Minamisawa, K., Uchiumi, T., Sasamoto, S., Watanabe, A., Idesawa, K., Iriguchi, M., Kawashima, K., *et al.* (2002) *DNA Res.* **9,** 189–197.
12. Wu, M., Sun, L. V., Vamathevan, J., Riegler, M., Deboy, R., Brownlie, J. C., McGraw, E. A., Martin, W., Esser, C., Ahmadinejad, N., *et al.* (2004) *PLoS Biol.* **2,** 327–341.
13. Gray, M., Burger, G. & Lang, B. F. (1999) *Science* **283,** 1476–1481.
14. Karlberg, O. & Andersson, S. G. E. (2003) *Nat. Rev. Genet.* **4,** 391–397.
15. Gabaldon, T. & Huynen, M. A. (2003) *Science* **301,** 609.
16. Karlberg, E. O., Canbäck, B., Kurland, C. G. & Andersson, S. G. E. (2000) *Yeast* **17,** 170–187.
17. Tamas, I., Klasson, L., Canbäck, B., Näslund, A. K., Eriksson, A.-S., Wernegreen, J. J., Sandström, J. P., Moran, N. A. & Andersson, S. G. E. (2002) *Science* **296,** 2376–2379.
18. Snel, B., Bork, P. & Huynen, M. (2002) *Genome Res.* **12,** 17–25.
19. Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. (2003) *BMC Evol. Biol.* **3,** 2.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Liplman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
21. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
22. Guindon, S. & Gascuel, O. (2003) *Syst. Biol.* **52,** 696–704.
23. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997) *Science* **278,** 631–637.
24. Swofford, D. L. (1998) Phylogenetic Analysis Using Parsimony (PAUP) (Sinauer, Sunderland, MA), Version 4.0b10.
25. Nimwegen, E. (2003) *Trends Genet.* **19,** 479–484.
26. Konstantinidis, K. T. & Tiedje, J. M. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 3160–3165.
27. Klasson, L. & Andersson, S. G. E. (2004) *Trends Microbiol.* **12,** 37–43.
28. Koonin, E. V. (2000) *Annu. Rev. Genomics Hum. Genet.* **1,** 99–116.
29. Kobayashi, K. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 4678–4683.
30. Olsen, G. J., Woese, C. R. & Overbeek, R. (1994) *J. Bacteriol.* **176,** 1–6.
31. Viale, A. & Arakaki, A. K. (1994) *FEBS Lett.* **341,** 146–151.
32. Emelyanov, V. (2003) *Arch. Biochem. Biophys.* **420,** 130–141.

Evolution Profonde et Phylogénie

Durant cette thèse je me suis intéressé à l'évolution profonde du vivant, depuis le dernier ancêtre commun universel (LUCA) jusqu'aux ancêtres des trois grands royaumes, les Archées, les Bactéries et les Eucaryotes. J'ai notamment cherché à placer quelques organismes dans l'arbre de la vie, tels que la bactérie *Aquifex aeolicus* et l'archée *Cenarchaeum symbiosum*, et j'ai également étudié l'évolution des températures de croissance il y a plusieurs milliards d'années. Pour ce faire, j'ai développé des algorithmes afin de reconstruire l'évolution de séquences géniques, puis j'ai utilisé ces séquences pour prédire les températures optimales de croissances d'organismes aujourd'hui éteints. Mes collègues et moi-même estimons que LUCA ne vivait pas à très haute température, mais que ses directs descendants les ancêtres des Bactéries et du groupe comprenant les Archées et les Eucaryotes vivaient dans des environnements plus chauds. Cela signifie que les deux lignées venant de LUCA ont subi le même type d'évolution en parallèle, qui pourrait avoir été causée par une seule et même pression de sélection. Cette pression pourrait être le résultat d'un intense bombardement météoritique il y a 3.8 milliards d'années, et avoir été accompagnée d'un changement depuis un génome à ARN pour LUCA vers des génomes à ADN pour ses descendants. Ensuite, dans la lignée des Bactéries, les températures optimales de croissance ont chuté, ce qui pourrait correspondre à l'évolution de la température des océans au cours des 3.5 derniers milliards d'années.

Early Evolution and Phylogeny

During this thesis, I studied the early evolution of life, from the Last Universal Common Ancestor (LUCA) to the ancestors of the three kingdoms, Archaea, Bacteria and Eukarya. Notably, I have attempted to place a few organisms in the tree of life, namely the bacteria *Aquifex aeolicus* and the archaea *Cenarchaeum symbiosum*, and I also studied the evolution of optimal growth temperatures over the last four billion years. To this end, I developed algorithms to reconstruct ancestral gene sequences, and used these sequences to predict the optimal growth temperatures of now-extinct organisms. My colleagues and I estimate that LUCA did not live in a very hot environment, but that its descendants the ancestors of Bacteria and of the group containing Archaea and Eukarya both lived at higher temperatures. This implies that the two lineages descending from LUCA underwent the same kind of evolution in parallel, perhaps caused by the same unique selection pressure. This pressure may have resulted from an intense meteoritic bombardment 3.8 billion years ago, and have been accompanied by the transition from an RNA genome in LUCA to DNA genomes in its descendants. Subsequently in the bacterial lineage, optimal growth temperature dropped, which may correspond to the evolution of oceanic temperatures in the last 3.5 billion years.