Bioinformatische Analysen zur Rekonstruktion tiefer phylogenomischer Stammbäume und Netzwerke am Beispiel früher evolutionärer Ereignisse

I n a u g u r a l - D i s s e r t a t i o n

zur Erlangung des Doktorgrades der

Mathematisch-Naturwissenschaftlichen Fakultät der

Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Mayo Röttger

aus Wolfenbüttel

Mai 2013

Aus dem Institut für Molekulare Evolution
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. William Martin
Korreferent: Prof. Dr. Martin Lercher

Tag der mündlichen Prüfung: 19. Juni 2013

Im Laufe dieser Arbeit wurden mit Zustimmung des Betreuers folgende Beiträge veröffentlicht:

Publikationen in Fachzeitschriften

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 25:748-761.

Roettger M, Martin W, Dagan T. 2009. A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. Mol Biol Evol. 26:1931-1939.

Dagan T***, Roettger M*****, Bryant D, Martin W**. 2010. Genome networks root the tree of life between prokaryotic domains. Genome Biol Evol. 2:379-392.

Stucken K, Ilhan J, Roettger M, Dagan T, Martin W. 2012. Transformation and conjugal transfer of foreign genes into the filamentous multicellular cyanobacteria (subsection v) *Fischerella* and *Chlorogloeopsis*. Curr Microbiol. 65:552-560.

Martin WF, Roettger M, Kloesges T, Thiergart T, Woehle C, Gould S, Dagan T. 2012. Modern endosymbiotic theory: Getting lateral gene transfer into the equation. J Endocytobiosis Cell Res. 23:1-5.

Dagan T***, Roettger M*****, Stucken K, Landan G, Koch R, Major P, Gould SB, Goremykin VV, Rippka R, Tandeau de Marsac N, Gugger M, Lockhart PJ, Allen JF, Brune I, Maus I, Pühler A, Martin WF**. 2012. Genomes of stigonematalean cyanobacteria (Subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol Evol. in press.

* Der Beitrag beider Autoren ist gleichwertig.

Tagungsbeiträge

Roettger M, Dagan T, Martin W. Inference of lateral gene transfer from phylogenetic trees is sensitive to multiple sequence alignment characteristics. SMBE Annual Meeting 2008. Barcelona, Spanien. Posterpräsentation

Roettger M, Dagan T, Martin W. Sequence alignment properties predict with 80 % accuracy the identification of lateral gene transfer events from discordant branches. SFB-TR1 Annual Meeting 2008, Martinsried, Deutschland. Posterpräsentation

Kürten A, Röttger M, Ahmadinejad N, Dagan T, Martin W. The role of symbiosis in forging eukaryotic genomes: Phylogenomics in the post genomic era. SFB-TR1 Annual Meeting 2009, Marburg, Deutschland. Posterpräsentation

Roettger M*, Dagan T*, Kleyner L, Bryant D, Martin W. The root of the tree of life is between archaebacteria and eubacteria. SMBE Annual Meeting 2009. University of Iowa, Vereinigte Staaten von Amerika. Posterpräsentation

Roettger M*, Dagan T*, Bryant D, Martin W. Genome networks root the tree of life between prokaryotic domains. Bertinoro Computational Biology 2010. University Residential Center, Bertinoro, Italien. Vortrag

Roettger M*, Dagan T*, Bryant D, Martin W. Genome networks root the tree of life between prokaryotic domains. SMBE Annual Meeting 2010. Lyon, Frankreich. Posterpräsentation

Roettger M, Martin W, Dagan T. Genes of cyanobacterial origin in plant nuclear genomes point to a filamentous true-branching heterocyst-forming plastid ancestor. The British Phycological Society winter meeting 2011. Cardiff, Vereinigtes Königreich. Vortrag

Roettger M*, Dagan T*, Stucken K, Landan G, Koch R, Major P, Gould SB, Goremykin VV, Rippka R, Tandeau de Marsac N, Gugger M, Lockhart PJ, Allen JF, Brune I, Maus I, Pühler A, Martin WF. Evidence for the origin of photosynthesis and plastids in large cyanobacterial genomes. SFB TR1 Annual Meeting 2012. München, Deutschland. Posterpräsentation

* Der Beitrag beider Autoren ist gleichwertig.

Inhaltsverzeichnis

1 Zusammenfassung

Phylogenetische und phylogenomische Analysen werden durch eine Vielzahl an Fehlerquellen beeinflusst, die zu Rekonstruktionsartefakten und Fehlinterpretation der tatsächlich stattgefundenen evolutionären Ereignisse führen können. Die betroffenen Schritte sind: Clusteranalyse, anschließendes Alignment orthologer Sequenzen, sowie die Rekonstruktion der Stammbäume.

Ein Beispiel, welches die inhärenten Probleme bei der Aufklärung tiefster Divergenzen in Stammbäumen verdeutlicht ist die Frage nach der ältesten Verzweigung in der Evolutionsgeschichte der Prokaryoten. Untersuchungen zur Bestimmung der Position einer Wurzel innerhalb einer prokaryotischen Phylogenie kamen bisher zu sehr widersprüchlichen Ergebnissen. In der vorliegenden Arbeit wurde ein verbreitetes Verfahren zur Gruppierung orthologer Sequenzen dazu verwendet, Proteinfamilien unterschiedlicher Altersgruppen zu generieren. Die aus dem Vergleich dieser Gruppen ermittelten Aufteilungsinformationen wurden zur Rekonstruktion phylogenetischer Netzwerke verwendet, auf deren Basis die Wurzel im Stammbaum der Prokaryoten zuverlässig zwischen Archaebakterien und Eubakterien, als die beiden ältesten Gruppen, bestimmt werden konnte. Fehleranfällige Schritte, wie die Rekonstruktion von Sequenzalignments und Stammbäumen, wurden dabei vermieden. Außerdem wurde gezeigt, dass es für die betrachteten Sequenzen nur einige wenige prokaryotische Gruppen mit erhöhten relativen Evolutionsraten gab, die Archaebakterien jedoch, entgegen der Aussagen in anderen Studien, nicht zu diesen gehörten.

Auf der Stufe des Sequenzalignments können Artefakte statistisch hinreichend unterstützte und dennoch falsch rekonstruierte Stammbäume generieren. Unstimmige Äste zwischen Genbäumen und einem Referenzstammbaum der betrachteten Organismen werden häufig als Anzeichen lateralen Gentransfers gedeutet. Analysen im Rahmen dieser Arbeit deuteten darauf hin, dass vor allem problematische Alignments Unstimmigkeiten als Artefakte bei der phylogenetischen Rekonstrukti-

on generieren. Untersuchungen anhand der Ergebnisse einer Studie zur Ableitung lateraler Gentransferereignisse, auf der Basis orthologer Genfamilien aus 144 prokaryotischen Genomen, zeigten, dass Veränderungen in den Eigenschaften der zugrundeliegenden Alignments die Anfälligkeit hierfür begünstigten und damit eine korrekte Ableitung der Stammbäume verhinderten. Allein mit Hilfe dieser Unterschiede in den Eigenschaften war es möglich, mit 80 prozentiger Sicherheit zu bestimmen, ob ein bestimmtes Genalignment hinreichend unterstützte unstimmige Äste generieren würde oder nicht, unabhängig von einem eventuell vorhandenen phylogenetischen Signal in den entsprechenden Sequenzen.

Die Bedeutung verlässlich erstellter Sequenzalignments zur Vermeidung von Artefakten in phylogenomischen Analysen wird noch einmal unterstrichen, wenn die Ableitung endosymbiotischer Gentransferereignisse vom Vorfahren der Plastiden in Algen- und Pflanzengenomen betrachtet wird. In Abhängigkeit der Alignmentverlässlichkeit konnte ein großer Anteil von Genen endosymbiotischen Ursprungs in den untersuchten photosynthetischen Eukaryoten ermittelt werden. Die Suche nach einem Cyanobakterium mit der zu einem hypothetischen Vorfahren der Plastiden ähnlichsten Gensammlung ergab, dass eine Phylogenie auf der Basis universell verteilter kernkodierter Proteine cyanobakteriellen Ursprungs und ihrer cyanobakteriellen Homologe kein verlässliches Ergebnis zulässt. Die erzeugten Stammbäume sowie die zugrundeliegenden Alignments enthielten Anzeichen für wohlbekannte Artefakte in der Rekonstruktion von Stammbäumen. Aus diesem Grund wurden alternativ die Verteilungsmuster und die mittlere Sequenzidentität der Algen- und Pflanzengene cyanobakteriellen Ursprungs in cyanobakteriellen Genomen untersucht. Hierbei konnte eine hohe Ähnlichkeit zu den Gensammlungen der heutigen Untergruppen IV und V festgestellt werden. Im Gegensatz zu allen anderen Cyanobakterien zeichnen sich deren Vertreter vor allem durch ihre zahlreichen Möglichkeiten der Zelldifferenzierung und der Möglichkeit molekularen Stickstoff in dafür spezialisierten Zellen zu binden, als eine sehr hoch entwickelte Gruppe aus. Sequenzanalysen ergaben, dass der komplexe filamentöse Phänotyp mit echten Verzweigungen der Untergruppe V vermutlich nur auf wenige Gene zurückzuführen ist, welche Einfluss auf Komponenten der Zellteilung ausüben. Aus universellen Proteinfamilien, die in allen untersuchten Genomen jeweils einfach kodiert vorlagen, wurde ein Stammbaum der Cyanobakterien erstellt. Dieser lässt einen Ursprung der oxygenen Photosynthese im Süßwasser vermuten und legt außerdem einen gemeinsamen Ursprung aller filamentösen Stämme nahe. Ein Netzwerk cyanobakteriellen Gentransfers auf Basis dieses Stammbaumes konn-

te zeigen, dass lateraler Gentransfer ein sehr häufiges Ereignis in der Evolution der Cyanobakterien darstellt, dabei jedoch selten viele Gene gemeinsam übertragen wurden. Die Verzerrung der cyanobakteriellen Phylogenie durch diesen verbreiteten Transfermechanismus konnte jedoch ausgeschlossen werden.

2 Summary

Phylogenetic and phylogenomic analyses are prone to copious sources of errors which can lead to reconstruction artifacts and misinterpretation of the true history of the evolutionary events that sequences in question have undergone. Affected operations are mainly cluster analysis, alignment of orthologous sequences, and the reconstruction of evolutionary trees.

One example, which illustrates inherent problems in solving deep divergences in phylogenetic trees is the question concerning the most ancient split in the evolutionary history of prokaryotes. Up until now, studying the position of a root in the prokaryotic phylogeny lead to very conflicting results. We were able to use a familiar method for clustering orthologous sequences to generate protein families of different age groups. The genome split information derived from comparisons between these groups were used to reconstruct phylogenetic networks. On basis of these networks, the root in the evolutionary tree of prokaryotes could be robustly estimated between archaebacteria and eubacteria as being the oldest prokaryotic groups. Therefore, error-prone steps like reconstruction of sequence alignments and phylogenetic trees were avoided. Moreover, elevated evolutionary rates were shown to exist only in a few prokaryotic lineages, archaebacteria being none of these as had been consistently suggested before.

At the stage of sequence alignment, artifacts can generate statistically well supported yet erroneous reconstructed trees. Conflicting branches between gene trees and a reference phylogeny of the analyzed organisms are often interpreted as indicators of lateral gene transfers. Analyses under this study indicated that majorly problematic alignments are generating these conflicts as artifacts in the phylogenetic reconstruction. In the analyzed dataset, changes in alignment properties supported this vulnerability and therefore impeded reliable reconstruction of trees. Only by the observed differences in properties it was possible to predict with 80 percent certainty, if a particular alignment would generate sufficiently supported

conflicting branches or not, disregarding the potentially existing phylogenetic signal in the respective sequences.

The importance of creating reliable sequence alignments to prevent artifacts in phylogenomic analyses is emphasized by the reconstruction of endosymbiotic gene transfer events from the ancestor of plastids into genomes of algae and plants. In dependance on alignment reliability, a hight proportion of genes of endosymbiotic origin could be estimated in the analyzed photosynthetic eukaryotic genomes. The attempt to identify the collection of genes with the highest similarity to that of a hypothetical ancestor of plastids among recent cyanobacterial genomes, based on a phylogeny of universally distributed nuclear encoded proteins of cyanobacterial origin and their cyanobacterial homologs, showed that this method can hardly be relied on. Reconstructed trees and underlying alignments proved to contain indicators of well known tree building artifacts. Hence, presence / absence patterns and average sequence identity of genes of cyanobacterial origin from algae and plants in cyanobacterial genomes were analyzed. Highest similarity could be observed with the collection of genes present in recent subsections IV and V.

From all other cyanobacteria, these representatives are distinguished by numerous capabilities in cell differentiation and the ability to fix molecular nitrogen in specialized cells as a highly evolved group of cyanobacteria. Analyzes revealed that the complex true-branching filamentous phenotype of subsection V members is presumably referable to a few genes which influence components of cell division. Universal protein families encoded in all studied genomes as single-copy genes, were used to reconstruct the cyanobacterial species tree. It supports the origin of oxygenic photosynthesis in fresh water environment and furthermore suggests a common ancestry of all filamentous strains. A cyanobacterial gene transfer network based on this backbone phylogeny marked lateral gene transfer as being a very frequent event in the evolution of cyanobacteria, however the coupled transfer of many genes was rare. A biased cyanobacterial species tree due to this prevailing transfer mechanism could be rejected.

3 Einleitung

3.1 Stammbäume früher und heute

Seit jeher beschäftigten sich die Menschen mit der Frage nach ihrer Abstammung. Ab dem 18. Jahrhundert warf vor allem die auf dem von Carl von Linné entwickelten hierarchischen System basierende Beschreibung zahlreicher neuer Pflanzen und Tierarten verstärkt die Frage nach der Herkunft und den Verwandtschaftsbeziehungen aller Lebewesen auf. Bevor molekulare Sequenzdaten von Organismen zur Verfügung standen, gab es bereits Versuche, evolutionäre Stammbäume von den Lebewesen abzuleiten und darzustellen. In den Anfängen dieser Disziplin war es dabei jedoch nur möglich, morphologische oder embryologische Merkmale unter den seinerzeit entdeckten und untersuchten Organismen zu bewerten und in einen verwandtschaftlichen Kontext einzuordnen. Die ersten Bäume spiegelten jedoch nicht die Beziehungen zwischen den Taxa wider, sondern stellten vielmehr Klassifizierungen durch Verbindungshierarchien oder Klammerungen dar und waren dadurch noch keine echten phylogenetischen Bäume (Tassy, 2010). Lamarcks Tafel der Entstehung der Tiere (Lamarck, 1809) sowie Barbancois detailreichere Version dieses Baumes (Barbançois, 1816) beschreiben die direkte Vorstufe phylogenetischer Bäume (Tassy, 2010). Charles Darwin skizzierte in seinen Aufzeichnungsbüchern 1837 den als „*I think*"-Baum bekanntgewordenen Entwurf zur Beschreibung der evolutionären Verwandtschaft zwischen den Lebewesen. Die einzige Abbildung in seinem bekannten Werk „*The Origin of species*"zeigt einen durch natürliche Variation und Selektion hervorgegangenen theoretischen phylogenetischen Baum (Darwin, 1859). Ernst Haeckel beschäftigte sich ebenfalls mit Stammbäumen. Abbildungen in seinen Büchern zeigten einen monophyletischen Stammbaum der Organismen (Haeckel, 1866) und die Abstammung des Menschen in der Phylogenie der damals bekannten Gruppen der Protozoen, Metazoen, Wirbeltiere und Säugetiere (Haeckel, 1874).

Erst etwa ein Jahrhundert später wurde mit der Entdeckung der DNA und ihrer Struktur als Träger der Erbinformationen aller Lebewesen (Watson und Crick, 1953) und Untersuchungen bezüglich der am besten geeigneten Moleküle als Basis für die Rekonstruktion molekularer Stammbäume (Zuckerkandl und Pauling, 1965) die Erforschung verwandtschaftlicher Beziehungen aufgrund von Gemeinsamkeiten und Unterschieden in ihren vererbten Molekülen selbst möglich. Neue Forschungen auf diesem Bereich waren vor allem auf dem Gebiet der Mikrobiologie bahnbrechend bei der Erforschung einer Systematik prokaryotischer Organismen. Für Bakterien, Algen und Protozoen war die bis zu diesem Zeitpunkt für Tiere und Pflanzen übliche Analyse morphologischer Merkmale unter Umständen problematisch, da sie teilweise keine komplexen intrazellulären Strukturen aufwiesen und sich morphologisch betrachtet nicht stark genug voneinander unterschieden (Graur und Li, 2000). Die Entwicklung der Proteinsequenzierungsmethoden und die Möglichkeit der schnellen Sequenzierung von Nukleinsäuren (Sanger, Nicklen und Coulson, 1977) ließen die Anzahl der verfügbaren Vergleichssequenzen stetig anwachsen und hatten einen großen Einfluss auf die molekulare Phylogenetik (Graur und Li, 2000). Es dauerte nicht lange bis Woese und Fox (1977) die Struktur der prokaryotischen Domäne mit der grundlegenden Aufteilung in Archaebakterien und Eubakterien auf der Basis ribosomaler RNA nachwiesen und damit die Einteilung aller Lebewesen in die drei Hauptreiche der Archaebakterien, Eubakterien und Eukaryoten ableiteten. Über die Verwandtschaftsbeziehungen dieser Gruppen im Stammbaum des Lebens gab es jedoch bis heute sehr kontroverse Hypothesen (O'Malley, Martin und Dupré, 2010).

Fortschritte in der automatischen Sequenzierungstechnologie sowie in den Computer-gestützten Berechnungsmethoden zur Assemblierung hunderttausender 300 bp bis 500 bp größer Stücke komplementärer DNA (Adams *et al.*, 1992, 1991) ermöglichten schließlich die schnelle, verlässliche und kosteneffektive automatische Sequenzierung kompletter Genome unabhängig von bereits existierenden Genomkarten mit Hilfe der sogenannten Shotgun-Sequenzierung (Fleischmann *et al.*, 1995). Dieser Zeitpunkt läutete eine neue Ära ein, in der Vergleiche zwischen kompletten Genomen zu einer unverzichtbaren Komponente der Aufklärung und des Verständnisses der Vielfalt biologischer Phänomene wurden. Vollständig sequenzierte Genome ermöglichten damit erstmals die Deliniearisierung des gesamten Netzwerks der Beziehungen zwischen Genen unterschiedlicher Genome (Tatusov, Koonin und Lipman, 1997).

3.2 Phylogenetische Rekonstruktion und inhärente Fehlerquellen

3.2.1 Gruppierung homologer Sequenzen

Das exponentiell steigende Angebot an Sequenzen der verschiedensten Organismen in den Datenbanken des angebrochenen Genomzeitalters erforderte die Ablösung der mehr oder weniger willkürlichen Gruppierung von Genen nach ihrer Ähnlichkeit durch vollständige und konsistente Gruppierungsverfahren für Gene, die aus einem gemeinsamen Vorfahren hervorgegangen waren (Tatusov, Koonin und Lipman, 1997). Die auf jedes Artentstehungsereignis folgende Divergenz der molekularen Sequenzen erzeugt eine Spur welche die Rekonstruktion von Stammesverzweigungen (Kladogenese) ermöglicht solange man ausschließlich zueinander orthologe (Fitch, 1970) Gene betrachtet (Fitch, 1970, 1995). Soll also der Stammbaum verschiedener Organismen auf der Basis eines Gens rekonstruiert werden, müssen zunächst die orthologen Sequenzen dieses Gens beziehungsweise des davon abgeleiteten Proteins aus den untersuchten Genomen identifiziert und in Gruppen zusammengefasst werden.

Es gibt verschiedene Verfahren, die der Identifikation orthologer Gene oder Genfamilien dienen, bei denen Paraloge, also Sequenzen deren Homologie eine Folge der Genduplikation ist (Fitch, 1970), demnach idealerweise ausgeschlossen werden sollten. Einige Methoden laufen nicht vollständig automatisch ab und erfordern die manuelle Betreuung (Tatusov *et al.*, 2003, 2001), andere sind vollständig automatisiert (Jensen *et al.*, 2008; Li, Stoeckert und Roos, 2003; Remm, Storm und Sonnhammer, 2001), benötigen deshalb jedoch sehr viel aufwendigere Algorithmen (Altenhoff und Dessimoz, 2009).

Roth, Gonnet und Dessimoz (2008) widmen sich mit ihrem Algorithmus OMA dem Problem der Ableitung von pseudo-orthologen Sequenzen, die durch differentiellen Verlust von Genen in einigen Genomen auftreten können (Altenhoff und Dessimoz, 2009). Weitere Beispiele für die Identifikation orthologer Sequenzen sind die im Ensembl-Projekt zur Annotation von Chordaten-Genomen implementierte Ableitung orthologer Sequenzfamilien durch Rekonstruktion und Angleichung von Genbäumen und Speziesbäumen (Hubbard *et al.*, 2007), sowie der paarweise Genvergleich in Kombination mit einem Leitbaum (engl. *guide tree*) und der konservierten Gennachbarschaft (Coordinators, 2013) wie sie in der Datenbank des *National Center for Biotechnology Information* (NCBI) angewendet wird (Altenhoff und Dessimoz, 2009). Ein sehr einfaches und dennoch effizientes Verfahren (Wolf

und Koonin, 2012), ist die Gruppierung von Paaren benachbarter bidirektionaler bester Treffer (Overbeek *et al.*, 1999).

Das TRIBE-MCL-Verfahren

Ein anderes Verfahren zur Darstellung orthologer Gruppen ist TRIBE-MCL (Enright, van Dongen und Ouzounis, 2002) und verwendet den sogenannten Markov-Gruppierungs-Algorithmus (engl. *clustering algorithm*) zur Gruppierung von Graphen durch Fluss-Simulation van Dongen (2000). Bei diesem Verfahren wird zunächst eine Matrix mit den Ähnlichkeiten der zu gruppierenden Sequenzen erstellt. Diese werden standardmäßig mit Hilfe einer BLAST-Suche (Altschul *et al.*, 1997) ermittelt, bei der jede Sequenz mit jeder anderen verglichen wird und die Trefferpaare mit ihren lokalen Sequenzidentitäten berechnet werden. Zur Erzeugung orthologer Sequenzfamilien kann das konservativere Verfahren der reziproken BLAST-Suche (Tatusov, Koonin und Lipman, 1997) mit anschließender Berechnung der Sequenzidentitäten aus einem paarweisen globalen Alignment, beispielsweise dem Needleman-Wunsch-Algorithmus, angewendet werden (Needleman und Wunsch, 1970).

TRIBE-MCL wurde vor allem entwickelt, um Probleme bei der Erstellung von Proteinfamilien aufgrund komplexer Multidomänenstrukturen zu lösen, die vor allem für große eukaryotische Datensätze auftreten (Enright, van Dongen und Ouzounis, 2002). Hierbei ist die Matrix ein Graph mit den Verbindungen zwischen den Sequenzen und einer Gewichtung der Äste, die den Ähnlichkeitsbeziehungen zwischen den Knoten entspricht. Diese werden in Verbindungswahrscheinlichkeiten überführt und die Matrix anschließend durch iterative Durchläufe von Matrix-Expansion durch Multiplikation sowie Matrix-Dekomprimierung verändert. Am Ende ändert sich die Matrix nur noch wenig und beschreibt die Gruppierung in die einzelnen Sequenzfamilien (Enright, van Dongen und Ouzounis, 2002). Die Abbildung 3.1 verdeutlicht das Verfahren anhand einiger Sequenzen und ihrer Ähnlichkeitsmatrix. In jedem Schritt werden durch die angewendeten Matrixoperationen die Verbindungswahrscheinlichkeiten von Ästen in Bereichen mit hoher Dichte erhöht, während sie in Bereichen mit geringer Dichte reduziert werden. Die Gesamtzahl der Verbindungen wird dadurch schrittweise reduziert, bis sich schließlich die eigentlichen Sequenzfamilien heraus kristallisieren.

Durch die iterative Reduzierung der Anzahl der Verbindungen beziehungsweise der Verbindungswahrscheinlichkeiten zwischen den Knoten werden universellere

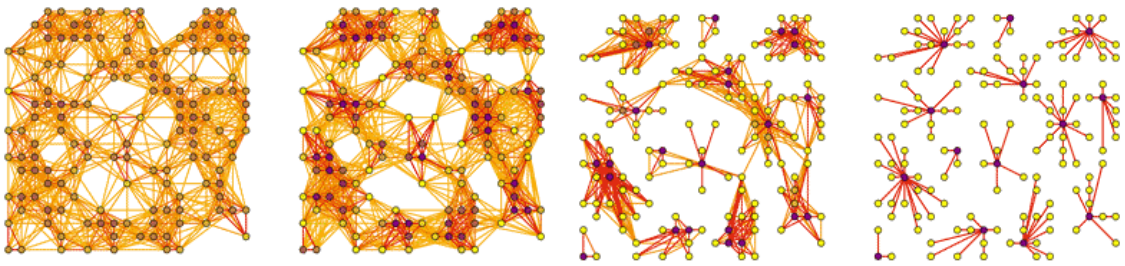


Abbildung 3.1: Gruppierungsverfahren des TRIBE-MCL-Algorithmus (Enright, van Dongen und Ouzounis, 2002).

Familien nach und nach in Sequenzfamilien mit spezifischeren Funktionen für eine kleinere Anzahl an Taxa aufgeteilt. Werden geringe paarweise Sequenzidentitäten zuvor ausgeschlossen, lassen sich die Familien während der Gruppierung tendenziell leichter in unterschiedliche spezifischere Familien aufteilen. Mit einer schrittweisen Erhöhung des Schwellenwertes für die paarweise Sequenzidentität wird es also möglich, den Zerfall von universelleren in spezifischere Familien zu beobachten und diese miteinander zu vergleichen. In Abschnitt 5.1 wird dieses Verfahren praktisch angewendet. Mit Genomaufteilungsmustern aus Vergleichen unterschiedlich gruppierter Sequenzfamilien unter aufsteigenden Schwellenwerten paarweiser Identität wurden phylogenetische Netzwerke rekonstruiert, welche die Ableitung der frühesten Stammesaufteilungen unter den Prokaryoten erlauben (Abschnitt 6.1).

Das auf der Graphentheorie beruhende TRIBE-MCL-Verfahren (Enright, van Dongen und Ouzounis, 2002) zur Gruppierung von Sequenzfamilien oder orthologen Gruppen stellt eine zuverlässige Lösung des Gruppierungsproblems, insbesondere für Proteinfamilien mit komplexen Multidomänenstrukturen wie sie typisch für eukaryotische Datensätze sind, bereit. Die Anwendung des beschriebenen Gruppierungsverfahrens kann bei der gemeinsamen Gruppierung vieler homologer Sequenzen sowohl prokaryotischen als auch eukaryotischen Ursprungs allerdings zu Problemen führen (Grünheit, 2010). Intradomänen-spezifische Verbindungen eukaryotischer Sequenzen auf der einen, und prokaryotischer Sequenzen auf der anderen Seite werden untereinander jeweils eine höhere Gewichtung aufweisen als die interdomänen-spezifischen Verbindungen zwischen Sequenzen von Prokaryoten und Eukaryoten. Am Ende des Gruppierungsverfahrens zerfallen dann interdomänen-spezifische Sequenzfamilien, welche gleichzeitig Orthologe von Prokaryoten und Eukaryoten beinhalten mit sehr hoher Wahrscheinlichkeit

jeweils in einzelne Teile. Ein Teil besteht dann aus den entsprechenden prokaryotischen Sequenzen während der andere nur den eukaryotischen Sequenzanteil beinhaltet. Diese Problematik kann dadurch vermieden werden, dass zunächst prokaryotische und eukaryotische Sequenzen separat von einander gruppiert werden. Erst anschließend werden die jeweiligen Proteinfamilien dann miteinander verbunden. Dieses Verfahren wurde in Abschnitt 5.3 angewendet.

3.2.2 Sequenzalignment

Zur Analyse molekularer Sequenzdaten wurde in der Anfangszeit ein bereits bekanntes hierarchisches Verfahren zur Erstellung von Gruppen (engl. *hierarchical clustering*) auf die beobachteten Unterschiede zwischen den betrachteten Sequenzen angewendet. Bei diesem Verfahren wurden schrittweise jeweils zwei Gruppen minimaler Distanz mit Hilfe von Ästen über einen hypothetischen gemeinsamen Vorfahren zu einer Gruppe höherer Ordnung verbunden. Die Distanz einer Gruppe zu einer anderen wird dabei in jedem Schritt über ihren arithmetischen mittleren Abstand neu berechnet (Sokal und Michener, 1958). Im ersten Schritt bestehen die Gruppen dabei lediglich aus den einzelnen zu vergleichenden taxonomischen Einheiten sowie einer Matrix, welche die paarweisen Unterschiede zwischen den Einheiten quantitativ beschreibt (Distanzmatrix).

Die Unterschiede zwischen den Sequenzen konnten dabei beispielsweise durch paarweise DNA-DNA-Hybridisierung ermittelt werden (Graur und Li, 2000). Die Idee war dabei, dass zwei Einzelstränge von verschiedenen Arten, welche über viele Wasserstoffbrücken ihrer komplementären Nukleinbasen interagieren, eine höhere Schmelztemperatur aufweisen als Sequenzen mit weniger Übereinstimmungen. Diese Unterschiede in den Schmelztemperaturen konnten damit direkt für das beschriebene Gruppierungsverfahren verwendet werden. Der inhärente Nachteile dieses Verfahrens waren die sehr aufwendige Kreuzhybridisierung und die Verwendung von Isotopen (Cho und Tiedje, 2001).

War dagegen die molekulare Sequenz der Nukleinsäuren bekannt oder standen Aminosäuresequenzen unterschiedlicher Proteine zur Verfügung, konnten die Unterschiede in der eigentlichen Sequenzabfolge direkt aus diesen quantifiziert werden. Entscheidend hierfür war allerdings, die homologen Bereiche der Sequenzen zu bestimmen. Dieses Sequenzalignment setzt als Bedingung voraus, dass die Sequenzen überhaupt mit einander vergleichbar sind, also einen gemeinsamen Ursprung haben (Feng und Doolittle, 1987). Das Alignieren von Sequenzen

beinhaltet die korrekte Identifikation homologer Nukleotide oder Aminosäuren und die Positionierung von Lücken, welche Insertionen oder Deletionen repräsentieren (Löytynoja und Goldman, 2008). Die homologen Bereiche von Sequenzen werden aneinander ausgerichtet, um Sequenzpositionen mit Schlüsselfunktionen zur Ableitung der evolutionären Geschichte zu ermitteln (Katoh *et al.*, 2002). Das Alignment ist die Voraussetzung für praktisch alle vergleichenden Sequenzanalysen und die phylogenetische Rekonstruktion und stellt damit eine fundamentale Funktion in der molekularen Biologie dar (Penn *et al.*, 2010).

Vom paarweisen zum multiplen Alignment

Obwohl der paarweise Vergleich nah verwandter Sequenzen bereits durch ausschließliche visuelle Betrachtung und Bewertung möglich war, führte dieses ermüdende und intuitive Verfahren bereits nach wenigen Jahren zur Entwicklung eines noch heute verwendeten Verfahrens zur Bestimmung eines optimalen globalen paarweisen Sequenzalignments, dem Needleman-Wunsch-Algorithmus (Needleman und Wunsch, 1970). Dieses Verfahren ermöglichte ein Alignment zweier Proteinsequenzen unter der Berücksichtigung von Insertions- und Deletionsereignissen, repräsentiert durch Lücken in einer der Sequenzen (engl. *gaps*) und zugehörigen Strafpunkten (engl. *gap penalty*). Diese erhöhen die Signifikanz der maximalen Übereinstimmung durch verringerte Gewichtung derjenigen Alignmentpfade, die viele Lücken in den Sequenzen erforderten (Needleman und Wunsch, 1970). Basis des Verfahrens war eine Austauschmatrix, die aufgrund der Klassifizierung der Aminosäuren, in übereinstimmende und nicht-übereinstimmende Paare sowie die Anzahl der jeweils korrespondierenden zugrunde liegenden Basen erstellt wurde (Needleman und Wunsch, 1970). Diese Matrix konnte später durch modernere Austauschmatrizen wie beispielsweise BLOSUM (Henikoff und Henikoff, 1992), JTT (Jones, Taylor und Thornton, 1992) oder WAG (Whelan und Goldman, 2001) ersetzt werden.

Eine spezielle Form des paarweisen Sequenzalignments entwickelten Smith und Waterman (1981). Ihr Algorithmus war in der Lage, ein optimales Abschnittspaar zweier Sequenzen (lokales Alignment) durch die Einführung von Deletionen und Insertionen beliebiger Länge zu finden (Smith und Waterman, 1981).

Beide Verfahren ließen sich prinzipiell auf höhere Dimensionen des Sequenzalignments verallgemeinern, in denen drei oder mehr Sequenzen gleichzeitig miteinander verglichen wurden (Smith, Waterman und Fitch, 1981), wie es spä-

ter die Programme MSA (Lipman, Altschul und Kececioglu, 1989) und DCA (Stoye, Moulton und Dress, 1997) zeigten. Problematisch wurde hierbei jedoch die Komplexität der Berechnung einer optimalen Lösung im multidimensionalen Sequenzraum.

Feng und Doolittle (1987) hatten schließlich die Idee, durch die iterative Anwendung des paarweisen Needleman und Wunsch Algorithmus ein multiples Alignment zu berechnen. Das Prinzip dieses Verfahrens beruht im wesentlichen auf dem schrittweisen paarweisen Alignment der Sequenzen anhand einer Distanzmatrix. Dabei werden in jedem Schritt die beiden Sequenzen oder Gruppen mit den geringsten Unterschieden paarweise aligniert. Die Positionen einmal eingeführter Lücken in die Sequenzen bleiben in folgenden Schritten unverändert (Feng und Doolittle, 1987). Auf dieser Methode beruht die Mehrzahl der heute verwendeten progressiven Alignmentverfahren, die teilweise die Sensitivität noch verbessern konnten ohne dabei Einbußen in Geschwindigkeit und Effizienz dieser Methode in Kauf zu nehmen (Edgar, 2004; Grasso und Lee, 2004; Notredame, Higgins und Heringa, 2000). CLUSTAL W ist ein weit verbreiteter Vertreter dieses Verfahrens (Dessimoz und Gil, 2010; Thompson, Higgins und Gibson, 1994).

Prinzipiell kann ein multiples Alignment dadurch verbessert werden, dass man es mit einem Strukturmodell vergleicht, das auf der Basis gleichwertiger struktureller Elemente der beteiligten Proteine der jeweiligen Proteinfamilie erstellt wurde (strukturelles Alignment), und daraus eine Bewertung der Güte eines jeweils möglichen Alignments ableitet. Die Schwierigkeit, ein Alignment auf diese Weise objektiv zu bewerten, lag zunächst vor allem in der begrenzten Anzahl solcher struktureller Alignments, welche als Modelle verwendet werden konnten (Gotoh, 1996). Mit der steigenden Anzahl dieser strukturellen Modelle für verschiedenste Proteinfamilien wurde es jedoch möglich, das multiple Alignment vor allem für divergente Proteinsequenzen mit einer iterativen Prozedur durch Verweise auf strukturelle Alignments gegenüber dem progressiven Alignment zu verbessern (Gotoh, 1996). Für viele Proteinfamilien fehlen heute leider immer noch strukturelle Informationen oder diese sind schwer anzuwenden (Notredame, Higgins und Heringa, 2000).

Die beiden Hauptprobleme des progressiven Alignments neben der Abhängigkeit der verwendeten Alignmentparameter beruhen auf der besonderen Strategie des Algorithmus (engl. *greedy*). Zum einen besteht das Problem des lokalen Minimums (Thompson, Higgins und Gibson, 1994). Des weiteren können einmal ins Alignment eingebaute Fehler sich auch noch in allen nach folgenden Alignment-

schritten negativ auswirken und dadurch vervielfältigen. Notredame, Higgins und Heringa (2000) versuchten diese Vervielfältigung der Fehler gerade in den ersten Schritten des Alignments zu minimieren, indem für das eigentliche multiple Alignment mit unterschiedlichen Programmen eine Bibliothek globaler und lokaler paarweiser Alignments der Sequenzen erstellt wird. In jedem Schritt des progressiven multiplen Alignments fließen dann jeweils Informationen darüber ein, wie alle Sequenzen untereinander alignieren, anstatt nur das aktuelle Paar zu betrachten (Notredame, Higgins und Heringa, 2000).

Ein weitererer interessanter Ansatz ist die Erzeugung eines Sequenzalignments basierend auf der schnellen Fourier-Transformation in Kombination mit einer iterativen Verbesserungsstrategie, wie er im Programm MAFFT implementiert wurde (Katoh *et al.*, 2002). Das Verfahren reduziert die benötigte Rechenzeit vor allem für sehr viele Sequenzen im Alignment. Dabei werden die einzelnen Aminosäuren einer Proteinsequenz zunächst in eine Sequenz aus Rauminhalts- und Polaritätswerten umgewandelt.

Das Programm MUSCLE (Edgar, 2004) verwendet ebenfalls eine progressive Alignmentstrategie, versucht jedoch die Rechenzeit für die zur Berechnung der Distanzmatrix nötigen paarweisen Alignments zu reduzieren. Hier werden die Distanzen der Sequenzen stattdessen aus übereinstimmenden k-Tupeln berechnet. Schließlich wird aus diesen der Leitbaum erzeugt. Da der Zeitaufwand für den ersten Schritt dadurch stark minimiert wird, werden die Laufzeiten inklusive des nachfolgenden progressiven Alignments sowie der Anwendung einer Verfeinerungsstrategie für ein Alignment tausender Sequenzen gegenüber t-Coffee und MAFFT ohne Einbußen bei der Präzision noch verbessert.

Bei Grasso und Lee (2004) wird das multiple Sequenzalignment nicht mehr in der üblichen linearen Repräsentation betrachtet, sondern durch einen gerichteten azyklischen Graphen ersetzt, welcher dadurch als partiell geordnet bezeichnet wird (engl. *partial order alignment*). Während beim progressiven Standardverfahren Sequenzen immer paarweise mit einer gemittelten Sequenz, also einem Profil, aligniert werden, erlaubt dieses Verfahren in jedem Schritt die Berücksichtigung der möglichen Alignments einer Sequenz mit allen möglichen homologen Rekombinationen der bereits im partiellen Alignment befindlichen Sequenzen. Dadurch gehen vor allem in komplexen Alignments in der Umgebung von Regionen mit vielen Lücken in den entsprechenden Sequenzen weniger Informationen verloren (Grasso und Lee, 2004).

Problematische Alignmentbereiche und phylogenetisches Signal

Die hauptsächliche Ursache für Fehler im Alignment von Sequenzen besteht darin, dass komplizierte evolutionäre Ereignisse wie Insertionen und Deletionen nicht getrennt als solche modelliert werden, sondern nur sehr vereinfacht durch das Einführen von Lücken in den entsprechenden Sequenzen basierend auf bestimmten Parametern dargestellt werden. Die Platzierung dieser Lücken ist es, die das automatisierte Sequenzalignment so problematisch macht (Higgins, Blackshields und Wallace, 2005; Kawakita *et al.*, 2003).

Eine verbreitete Methode, das Problem des fehlerhaften Alignments in Bereichen mit vielen Insertionen oder Deletionen zu umgehen, ist, diese Bereiche vor weiteren Analyseschritten zu entfernen. Problematisch ist dabei allerdings, dass gerade in diesen Bereichen auch phylogenetische Signale enthalten sind (Dessimoz und Gil, 2010). Darüber hinaus bleiben für konservierte Proteine auf diese Weise am Ende nur wenige und vorwiegend nicht informative Positionen übrig. Castresana (2000) schließt mit Hilfe des Programms GBLOCKS diese Bereiche aus, wobei der Verlust an informativen Positionen jedoch möglichst gering gehalten werden soll. Ein Ergebnis ihrer Arbeit war, dass in derartig behandelten Alignments die Aminosäurezusammensetzung der Sequenzen einheitlicher war und zum anderen die beobachteten Distanzen zwischen den Sequenzen abnahmen. Für basalere Äste in tiefen Phylogenien konnte das Verfahren jedoch keine wirkliche Verbesserung durch Erhöhung der Unterstützungswerte (engl. *branch-support*) liefern (Castresana, 2000). Einiges deutet darauf hin, dass gerade die Signale für die Auflösung tiefer Phylogenien in diesen problematischen Bereichen enthalten sind (Kawakita *et al.*, 2003) und sich die Entfernung von Bereichen mit Lücken nachteilig auswirkt (Dessimoz und Gil, 2010).

Alignmentfehler in schwer zu rekonstruierenden Bereichen mit vielen Insertionen und Deletionen sollten idealer Weise minimiert werden, um auf diese Weise den Informationsgehalt der Sequenzen nicht zu verfälschen und ihn in seiner Gesamtheit zu bewahren und nutzbar zu machen. Die traditionell verwendeten Alignmentalgorithmen beschränken sich auf schlichte Positionierung von Lückenzeichen als vereinfachte Modelle der Insertion und Deletion. Zwangsläufig führt dies zu einem Übermaß an Deletionen und Sequenzaustauschen einhergehend mit sehr wenigen Insertionen und der unplausiblen Rekonstruktion dieser Ereignisse (Löytynoja und Goldman, 2008). Die Ursache hierfür ist vor allem, dass in den meisten verwendeten progressiven Methoden Insertionen gegenüber Deletionen

systematisch benachteiligt werden, da für eine Insertion multiple Lückenstrafpunkte vergeben werden müssen, ganz im Gegensatz zu einer Deletion (Löytynoja und Goldman, 2005, 2008).

Am Beispiel eines traditionellen Sequenzalignments für ein HIV und SIV Hüll-Glykoprotein und zusätzlichen Sequenzsimulationen konnte gezeigt werden, dass herkömmliche Verfahren keine überzeugende Methode für die Evolution dieser Region liefern, da sie beim Alignieren der Regionen nahe beieinander liegender Insertionen (kollabierte Insertionen, engl. *collapsed insertions*) versagen (Löytynoja und Goldman, 2008). Das vorgestellte Alignmentverfahren verhindert das Verschmelzen von benachbarten jedoch unabhängigen Insertionsorten dadurch, dass Insertionsereignisse zunächst auch als solche markiert werden. Dadurch kann anschließend das Alignment dieser Sequenzstücke mit nicht-homologen Sequenzen verhindert werden. Stellt sich in den folgenden Schritten des progressiven Alignments jedoch heraus, dass es sich tatsächlich um Deletionsereignisse handelt, kann die Markierung für diese Sequenzstücke wieder aufgehoben werden.

Eine allgemeine Tendenz zu kompakten Alignments aus visuell ansprechend aneinander ausgerichteten übereinstimmenden Blöcken mit möglichst wenigen Lücken sollte gerade für nicht-kodierende DNA-Sequenzen oder Teile von Proteinsequenzen, in denen Insertionen und Deletionen weniger organisiert stattgefunden haben, hinterfragt werden (Higgins, Blackshields und Wallace, 2005).

Quantifizierung von Unsicherheiten als Folge des Alignmentverfahrens

Das Alignmentverfahren selbst kann bereits unabhängig von den eingestellten Parametern Artefakte erzeugen. Für bestimmte Sequenzbereiche ist das Alignment deshalb unter Umständen unzuverlässiger als für andere (Landan und Graur, 2007). Im progressiven Alignment wird üblicherweise in jedem Schritt ein optimales paarweises Alignment mit Hilfe des Verfahrens der dynamischen Programmierung (engl. *dynamic programming*) erzeugt. Gerade in Bereichen mit vielen Lücken sind oft alternative Alignmentpfade möglich, die ihrerseits ein ebenso optimales Alignment liefern. Vom jeweiligen Alignmentprogramm wird normalerweise nur eine dieser Möglichkeiten betrachtet und weiter verwendet. Diese Ambiguitäten können unterschiedlich auf das Alignment verteilt sein, nur kurze Bereiche betreffen oder sich sogar auf weite Teile erstrecken. Es ist naheliegend anzunehmen, dass weiterführende Analysen und die hieraus abgeleiteten Ergebnisse auf der Basis weniger verlässlich alignierter Bereiche ebenfalls zu unsicheren oder falschen

Ergebnissen führen können.

Landan und Graur (2007) stellten eine einfache Methode vor, welche die rasche Identifizierung und Quantifizierung dieser bis dahin vernachlässigten Unsicherheiten in multiplen Alignments und nachfolgender Analysen erlaubte, die sogenannte *Heads-or-tails*-Methode. Die Methode beruht auf der einfachen Idee, dass ein Alignment unabhängig von der anfänglichen Orientierung der Sequenzen sein sollte (5' oder 3', N-Terminus oder C-Terminus). Ein idealer Algorithmus sollte in jedem Fall das selbe Alignment erzeugen. In der Realität ist dies jedoch nicht der Fall: aus Unterschieden in den Alignments der jeweiligen Sequenzen in der originalen sowie in umgekehrter Orientierung können zum einen Unsicherheiten im Alignment quantifiziert werden, zum anderen erlaubt die Methode die Aufteilung eines Alignments in sichere und unsichere Bereiche. Mit Hilfe dieses Verfahrens konnte im Laufe dieser Arbeit beispielsweise analog zu Deusch *et al*. (2008) ein starker Zusammenhang zwischen der Qualität eines Alignments und des geschätzten Anteils endosymbiotischer Gentransferereignisse in Algen und Pflanzen festgestellt werden. Ignoriert man diesen Zusammenhang, werden diese Ereignisse sehr stark unterschätzt (Abschnitt 5.3, Abschnitt 6.3).

Penn *et al*. (2010) konnten mit einer weiteren Methode zur Quantifizierung positionsspezifischer Alignmentverlässlichkeit nachweisen, dass Unsicherheiten in den bei progressiven Alignmentverfahren verwendeten Leitbäumen die Hauptursache von Unsicherheiten im Sequenzalignment darstellen. In diesem Verfahren werden aus einem berechneten Basisalignment durch zufällige Auswahl von Spalten mit Zurücklegen mehrere neue Pseudoalignments erzeugt (engl. *bootstrapping*). Eine aus diesen Pseudoalignments generierte Menge abgeleiteter Leitbäume wird anschließend dazu verwendet, um alternative Alignments der ursprünglichen Sequenzen zu berechnen. Aus dem Vergleich des Basisalignments mit diesem Satz alternativer Alignments können dann sehr genau unsichere Positionen im multiplen Alignment identifiziert werden. Es konnte weiterhin gezeigt werden, dass ihr vorgestelltes Programm GUIDANCE bei der Vorhersage unverlässlicher Regionen im Alignment die *Heads-or-tails*-Methode (Landan und Graur, 2007) sogar noch in ihrer Leistung übertrifft.

3.2.3 Stammbäume

Standardverfahren zur phylogenetischen Rekonstruktion

Ab den 1970er Jahren wurden Methoden zur Ableitung phylogenetischer Bäume auf der Basis molekularer Sequenzdaten nach dem Prinzip der minimalen Evolution (engl. *minimum evolution*) beziehungsweise maximaler Parsimonie entwickelt. Beim Standardalgorithmus werden dabei alle für die betrachteten Sequenzen möglichen Topologien für die betrachteten Sequenzen untersucht, oder eine Teilmenge dieser, welche mit hoher Wahrscheinlichkeit der korrekten Phylogenie sehr ähnlich ist. Anschließend wird die Topologie ausgewählt, welche die wenigsten evolutionären Veränderungen benötigt um die beobachteten Sequenzunterschiede zu erklären (Saitou und Nei, 1987). Distanzbasierte Methoden wurden implementiert, welche das Problem des zeitintensiven Standardalgorithmus dadurch löste, dass der ermittelte Baum nicht zwangsläufig dem Baum minimaler Evolution entsprechen musste (Faith, 1985; Farris, 1972; Fitch, 1981; Sattath und Tversky, 1977; Tateno, Nei und Tajima, 1982). Die Effizienz den korrekten Baum zu erhalten war hier jedoch oft höher als bei der Anwendung des Standardalgorithmus (Saitou und Nei, 1987).

Ein anderer Ansatz, einen eindeutigen Baum unter dem Prinzip der minimalen Evolution zu berechnen die heute sehr verbreitete und effiziente Methode zur Verbindung von Nachbarn (engl. *neighbor-joining*, NJ) von Saitou und Nei (1987). Es werden jeweils Paare operativer taxonomischer Einheiten (engl. *operational taxonomic units*, OTUs) gesucht, welche beginnend mit einer Sterntopologie in jedem Schritt der Gruppierung der Einheiten die Gesamtlänge aller Äste minimiert (Saitou und Nei, 1987).

Ein Nachteil der Neighbor-Joining-Methode, beziehungsweise aller distanzbasierten minimalen Evolutionsmethoden, ist die Konvertierung der beobachteten Sequenzunterschiede in evolutionäre Distanzen. Hierbei werden die verfügbaren Daten komprimiert und die korrekte Berechnung der Distanzen für divergente Sequenzen wird in zunehmendem Maße von der Korrektur multipler Substitutionen abhängig (Holder und Lewis, 2003). Eine Möglichkeit, auf diese Überführung zu verzichten, wurde mit der auf Sequenzzeichen basierenden Parsimoniemethode zur phylogenetischen Rekonstruktion erreicht. Im Gegensatz zu den distanzbasierten Methoden beruht diese nicht mehr auf einer Matrix paarweiser Unterschiede. Stattdessen wird die evolutive phylogenetische Geschichte der Sequenzen jeweils auf einen Baum projiziert und jeder mögliche Baum bezüglich der minimal be-

nötigten Mutationen bewertet. Ein entscheidender Nachteil dieser Methode ist jedoch ihre starke Neigung zum Versagen bei konvergenter Evolution an langen Ästen und das häufige Auftreten gleich-parsimonischer (engl. *equally parsimonious*) Bäume (Holder und Lewis, 2003).

Ein weiteres sehr verbreitetes Verfahren zur Rekonstruktion phylogenetischer Bäume basiert auf der maximalen Wahrscheinlichkeit (engl. *maximum likelihood*, ML) (Felsenstein, 1973, 1981). ML berechnet für jeden möglichen Stammbaum zu vergleichender Sequenzen eine Wahrscheinlichkeit, die beschreibt, wie gut der entsprechende Baum die zu beobachtenden Unterschiede in den Sequenzen erklärt (Felsenstein, 1981). Der Baum mit der größten Wahrscheinlichkeit wird ausgewählt. Die Wahrscheinlichkeit berechnet sich dabei auf der Basis eines Evolutionsmodels, welches die relativen Wahrscheinlichkeiten der beobachtbaren Ereignisse beschreibt. Diese Modelle berücksichtigen auch die Wahrscheinlichkeiten unsichtbarer Ereignisse und können somit multiple Austausche an der gleichen Sequenzposition korrigieren (Holder und Lewis, 2003).

Parsimonie, sowie auch die Methode maximaler Wahrscheinlichkeit, sehen die Implementierung eines erschöpfenden (engl. *exhaustive*) Algorithmus vor, bei dem jeder mögliche Baum betrachtet wird. Jeder dieser Bäume erhält hierbei eine Bewertung, so dass sich am Ende die Phylogenie unter dem gewünschten Optimierungskriterium unter allen betrachteten Bäumen auswählen lässt. Da die Anzahl möglicher, verschiedener Bäume exponentiell mit der Anzahl der enthaltenen Sequenzen ansteigt, ist es spätestens ab 20 Sequenzen praktisch unmöglich eine vollständige Suche durchzuführen (Felsenstein, 1981; Graur und Li, 2000). Aus diesem Grund wurden heuristische Verfahren entwickelt, die ausgehend von einer Ausgangstopologie, zum Beispiel Neighbor-Joining, nur zu dieser ähnliche Bäume untersuchen (Graur und Li, 2000). Da das Aussortieren von Bäumen nach bestimmten Kriterien nur eine Abschätzung darstellen kann, besteht die Möglichkeit, dass die erstellte Phylogenie nicht dem optimalen Baum entspricht, der bei der vollständigen Suche gefunden würde (Felsenstein, 1981). Die generelle Güte eines heuristischen Verfahrens kann mit Hilfe von Sequenzsimulationen abgeschätzt werden, bei denen die tatsächliche Phylogenie der Sequenzen immer bekannt ist (Stoye, Evers und Meyer, 1998). Im Einzelfall lassen sich aus den aufgeführten Gründen jedoch keine Rückschlüsse darauf ziehen, ob der Algorithmus tatsächlich den optimalen korrekten Baum ermittelt hat oder der Baum nur eine Näherung ist.

Eine jüngere Entwicklung auf dem Gebiet der Rekonstruktion von Stammbäu-

men stellt die Bayes'sche Ableitung der Phylogenie dar (Huelsenbeck *et al*., 2001; Larget und Simon, 1999; Mau, Newton und Larget, 1999; Rannala und Yang, 1996). Die entscheidende Größe ist hierbei die *a posteriori*-Wahrscheinlichkeit, als Wahrscheinlichkeit des korrekt rekonstruierten Baumes. Im Bayes'schen Theorem wird diese aus der *a priori*-Wahrscheinlichkeit einer Phylogenie zusammen mit der Wahrscheinlichkeit die zu beobachtenden Daten für diesen Baum zu erhalten, und der *a priori*-Wahrscheinlichkeit für die gegebenen Daten berechnet (Huelsenbeck *et al*., 2001). Da die Berechnung der *a posteriori*-Wahrscheinlichkeit wegen ihrer Komplexität in der Praxis normalerweise nicht möglich ist, wird ein Markov-Ketten Monte Carlo (MCM)-Algorithmus verwendet, um diese abzuschätzen (Huelsenbeck *et al*., 2001). Der Grundgedanke hinter dieser Idee ist dabei, eine Markov-Kette mit den Parametern des statistischen Models als Zustandsraum und die Verteilung der *a posteriori*-Wahrscheinlichkeiten der Parameter als stationäre Wahrscheinlichkeitsverteilung zu erstellen (Huelsenbeck *et al*., 2001). Der MCM-Algorithmus besteht im wesentlichen aus zwei Schritten, wobei zunächst ein neuer Baum durch stochastische Veränderung des aktuellen Baumes vorgeschlagen wird. Dieser neue Baum wird dann mit einer bestimmten Wahrscheinlichkeit entweder beibehalten oder verworfen. Wird der Baum beibehalten, ist er die Grundlage weiterer Veränderungen. Eine Schwierigkeit bei dieser Methode ist, dass die verwendeten Markov-Ketten nicht immer in eine stationäre Verteilung konvergieren, weil die Kette beispielsweise zu lang wird oder der Mechanismus neue Bäume vorzuschlagen nicht optimal ausgelegt ist. In einer geeignet konstruierten und hinreichend lange gelaufenen Markov-Kette entspricht die *a posteriori*-Wahrscheinlichkeit eines Baumes genau dem Anteil, wie oft dieser Baum gefunden wurde (Huelsenbeck *et al*., 2001).

Von der Phylogenetik zur Phylogenomik

Phylogenomische Methoden zur Rekonstruktion von Stammbäumen stellen in den meisten Fällen einfache Erweiterungen phylogenetischer Standardmethoden dar, welche auf Datensätze einzelner Gene angewendet wurden (Philippe *et al*., 2011). Bei der Supermatrix-Methode werden beispielsweise die Alignments einzelner Gene in einem zusammenhängenden Alignment zusammengefügt und anschließend nach phylogenetischen Standardverfahren bearbeitet. Problematisch ist dabei allerdings, dass die einzelnen Gene mit unterschiedlichen Austauschraten evolvieren können. Die Methode benutzt hingegen nur ein einfaches Substitutionsmodell,

welches die Heterogenität zwischen Genen im evolutionären Prozess ignoriert und dadurch systematische Verzerrungen in den abgeleiteten Bäumen hervorrufen kann (Rannala und Yang, 2008). Dies muss in jedem Fall bei der anschließenden Analyse eines zusammengefügten Alignments berücksichtigt werden. Bei einer anderen Methode werden zunächst alle Stammbäume aus den Einzelalignments erstellt. Anschließend werden diese Einzelbäume dann mit Hilfe eines heuristischen Verfahrens in einer Konsensusphylogenie vereint (engl. *supertree*). Die *Supertree*-Methode neigt durch die unabhängige Bestimmung der jeweiligen Parameter für jedes einzelne Gen zur Überanpassung der Daten, wobei sich die Varianzen vergrößern. Des Weiteren fehlt eine statistische Basis für die verwendeten heuristischen Methoden zur Berechnung der Konsensusphylogenie (Rannala und Yang, 2008).

Die Verwandtschaft einzelner Gene oder Proteine kann sehr gut durch phylogenetische Bäume beschrieben werden. Mit der Sequenzierung vollständiger Genome und der Möglichkeit viele Gene einer Art und ihre Orthologe gemeinsam betrachten zu können entstanden jedoch neue Probleme. Es gibt evolutionäre Prozesse, welche sich prinzipiell nicht auf der Basis eines bifurzierenden Baumes beschreiben lassen. Diese sind Hybridisierung, endosymbiotischer Gentransfer, Verteilung von molekularen Merkmalen kurz nach einem Artentstehungsereignis (engl. *lineage-sorting*) und lateraler Gentransfer vor allem unter Prokaryoten durch Transduktion, Transformation oder Konjugation (Rannala und Yang, 2008; White *et al.*, 2007). In diesen Fällen können die herkömmlichen Methoden zur Bestimmung phylogenomischer Stammbäume versagen, da bezüglich der Artentstehungsereignisse verschiedene widersprüchliche Signale möglich sind und einzelne Genbäume vom Stammbaum der jeweils betrachteten Arten abweichen können (Bryant und Moulton, 2004; Fitch, 1997). Dieses Problem lässt sich vermeiden, wenn die Informationen großer Sammlungen von Einzelbäumen in einem phylogenetischen Netzwerk zusammengefasst werden. Die Erzeugung phylogenetischer Netzwerke kann durch die Zerlegung der Äste (engl. *split decomposition*) (Bandelt und Dress, 1992), *NeighborNet* (Bryant und Moulton, 2004) oder Konsensusnetzwerke (Holland, Delsuc und Moulton, 2005) erfolgen. Im Gegensatz zu einem phylogenomischen Konsensusbaum, in dem nur die Verzweigungen mit dem größten Anteil angezeigt werden können, können in Netzwerken auch diese widersprüchlichen Signale dargestellt werden (Bryant und Moulton, 2004; Fitch, 1997).

Allerdings können die beschriebenen Konflikte in den Genbäumen ebenfalls eine Folge von Zufallsprozessen und systematischen Fehlern in der phylogeneti-

schen Rekonstruktion (Rannala und Yang, 2008; Stiller, 2011). Parallele Evolution, Heterogenität des Models und Stichprobenfehler können die Bestimmung eines eindeutigen Baums erschweren (Bryant und Moulton, 2004). Die genannten Effekte reduzieren sich meist mit der Betrachtung einer größeren Menge an Positionen eines Gens (Rannala und Yang, 2008). Bei der Interpretation unstimmiger Stammbäume für verschiedene Gene müsste in diesem Zusammenhang daher stets kritisch geprüft werden, ob alternative Erklärungen wahrscheinlicher sind, bevor sie als Folge lateralen Gentransfers beurteilt werden. In Abschnitt 5.2 beziehungsweise Abschnitt 6.2 wird gezeigt, dass widersprüchliche Stammbaumtopologien eine Folge unverlässlich erstellter Alignments für problematische Sequenzen sein können.

Fehlerquellen bei der Rekonstruktion von Stammbäumen

Wie die vorhergehenden Schritte der Gruppierung orthologer Sequenzen und des Sequenzalignments ist auch die Ableitung phylogenetischer oder phylogenomischer Stammbäume anfällig für vielfältige Fehlerquellen. Einerseits können sich Fehler bei der Gruppierung orthologer Sequenzen hier stark negativ auswirken. Nur mit Hilfe orthologer Sequenzen können die zugrundeliegenden Stammverzweigungsereignisse korrekt abgeleitet werden (Fitch, 1970, 1995). Des Weiteren können sich Fehler im Alignment der untersuchten Sequenzen nachteilig auf die korrekte Rekonstruktion ihrer Phylogenie auswirken (Landan und Graur, 2007). In der vorliegenden Arbeit konnte beispielsweise ein großer Einfluss der Verlässlichkeit zugrunde liegender Alignments mit dem Anteil abgeleiteter Gentransferereignisse aus den von diesen abgeleiteten Stammbäumen festgestellt werden (Abschnitt 5.3, Abschnitt 6.3).

Zusätzlich zu diesen sich fortsetzenden Fehlern aus nötigen vorgeschalteten Schritten auf dem Weg zur Ableitung eines korrekten Stammbaums, sind die einzelnen phylogenetischen Rekonstruktionsmethoden mehr oder weniger stark von inhärenten Fehlerquellen betroffen. Dies soll im Folgenden näher erläutert werden. Als erstes sind hier die heuristischen Verfahren genannt. Es wird ein optimaler Stammbaum gesucht, der die evolutionäre Geschichte der Sequenzen korrekt widerspiegelt. Allerdings ist die Lösung dieses Problems durch die nötige Betrachtung aller möglichen Bäume nicht in polynomialer Zeit lösbar (NP-vollständig). Deshalb müssen zur Lösung von Problemen dieser Art heuristische Methoden angewendet werden. Es kann jedoch nicht garantiert werden, dass diese tatsächlich

den Baum finden, der dem globalen Optimum im Suchraum entspricht (Felsenstein, 1981; Penny, Hendy und Steel, 1992). Mit anderen Worten bedeutet dies, dass es unter Umständen eine Phylogenie gibt, die die beobachteten Sequenzdaten besser beschreiben würde mit der logischen Konsequenz, dass der ausgegebene Baum falsch ist.

Wie bereits bekannt ist, enthalten auch Regionen mit Lücken im Alignment ein phylogenetisches Signal (Dessimoz und Gil, 2010; Kawakita *et al.*, 2003). In den meisten Programmen zur phylogenetischen Rekonstruktion wird diese Informationen allerdings systematisch ignoriert (Lloyd und Calder, 1991; Penny, Hendy und Steel, 1992). Dabei ist die sehr einfache Nutzung des phylogenetischen Signals von Insertionen und Deletionen ein wichtiger Vorteil der Parsimoniemethode. Es konnte sogar gezeigt werden, dass dieses Verfahren für den speziellen Fall, dass Insertions- und Deletionsereignisse korrekt identifiziert wurden, sogar andere Verfahren bei der Rekonstruktion vor allem tiefer Phylogenien übertreffen kann (Dessimoz und Gil, 2010).

Grundlage für die fehlerfreie Rekonstruktion von Stammbäumen mit Distanzmethoden, maximaler Wahrscheinlichkeit oder nach Bayes ist jeweils ein Evolutionsmodell, das die Veränderungen der Sequenzen zuverlässig beschreibt (Abascal, Zardoya und Posada, 2005). Die Wahl eines passenden Modells ist entscheidend. Eine generelle Schwäche von Methoden maximaler Wahrscheinlichkeit liegt dabei in der hohen Unsicherheit der Interpretation der Sequenzzeichen, so dass es schwierig wird, jemals ein vertrauenswürdiges Wahrscheinlichkeitsmodel auszuwählen (Felsenstein, 1978). Wird ein unpassendes Modell verwendet, können multiple Substitutionen nicht mehr zuverlässig erkannt werden. Dies führt zur Verstärkung des nicht-phylogenetischen Signals und kann damit zu Artefakten bei der Rekonstruktion von Stammbäumen führen (Philippe *et al.*, 2011). Vor allem die unentdeckte oder ignorierte Variation der Evolutionsraten zwischen einzelnen Alignmentpositionen als bedeutender evolutionärer Prozess (Lopez, Casane und Philippe, 2002) in Verbindung mit Substitutionsratenvariation zwischen verschiedenen Stämmen (engl. *heterotachy*) führt zu falschen Ergebnissen bei der Anwendung Modell-basierter Verfahren zur Stammbaumrekonstruktion wie Distanzmatrixmethoden oder maximale Wahrscheinlichkeit (Yang, 1996). So gibt es Fälle in denen alle Standardmethoden der phylogenetischen Rekonstruktion versagen, falls Heterogenitäten zwischen den Evolutionsraten der untersuchten orthologen Sequenzen verschiedener Stämme vorhanden sind (Lockhart *et al.*, 2006).

Abascal, Zardoya und Posada (2005) bieten mit ihrem Programm ProtTest die Möglichkeit, das passendste Evolutionsmodel aus einer Auswahl an verschiedenen Modellen zu bestimmen. Natürlich kann es auch hier vorkommen, dass das am besten passende Model die reale Evolution der Sequenzen nicht optimal modelliert, jedoch kein besseres Model zur Verfügung steht. Es ist schwer abzuschätzen, in welchem Ausmaß die Konsistenz verschiedener phylogenetischer Rekonstruktionsmethoden durch Abweichungen des verwendeten Evolutionsmodels von den realen Vorgängen beeinflusst wird (Penny, Hendy und Steel, 1992). Beispielsweise können Unterschiede im GC-Gehalt der untersuchten Sequenzen fehlerhafte Gruppierungen verursachen (Penny, Hendy und Steel, 1992; Penny *et al*., 1990).

Beim Vergleich von nur vier DNA-Sequenzen ließen sich mit Hilfe von Simulationen die folgenden generellen Aussagen bezüglich der relativen Effizienz der vorgestellten Verfahren feststellen. Sowohl die gewichtete als auch die ungewichtete parsimonische Rekonstruktion einer Phylogenie sind generell weniger effizient als *Neighbor-Joining* oder die Methode maximaler Wahrscheinlichkeit. Des weiteren ist die Methode maximaler Wahrscheinlichkeit der *Neighbor-Joining*-Methode nur dann leicht überlegen, wenn alle Annahmen korrekt sind. Eine Verletzung der Grundannahmen beeinflusst eher die korrekte Bestimmung der Astlängen als die Ableitung der korrekten Topologie (Tateno, Takezaki und Nei, 1994). Sehr divergente Sequenzen verhalten sich in etwa wie zufällige Sequenzen und enthalten deshalb wenige phylogenetische Informationenen. In diesem Fall ist durch die Ungültigkeit der verwendeten Evolutionsmodelle jede Methode zur Bestimmung der Phylogenie dieser Sequenzen unzuverlässig (Nei, Takezaki und Sitnikova, 1995; Rzhetsky und Nei, 1995). Die Methode maximaler Wahrscheinlichkeit ist anfällig für Änderungen des Substitutionsmusters über den betrachteten evolutionären Zeitraum. Hier geht der Vorteil dieser Methode gegenüber anderen, die weniger strenge Annahmen über das Substitutionsmuster benötigen, verloren (Nei, 1996). Ausserdem wird die generelle Wahrscheinlichkeit, einen korrekten Baum zu erhalten, mit der Anzahl der betrachteten Sequenzen geringer (Nei, 1996).

Neben der Auswahl eines passenden Evolutionsmodells ist vor allem auch die Gruppierung langer Äste bei der Rekonstruktion einer Phylogenie problematisch (engl. *long-branch attraction artifacts*, LBA (Felsenstein, 1978)). Die zu untersuchenden Ereignisse liegen oft sehr weit zurück in der evolutionären Geschichte der Sequenzen, wodurch lange terminale Äste und damit einhergehende multiple Substitutionen an gleicher Position, also Homoplasien, unvermeidlich sind (Philippe *et al*., 2011). Hier werden die Sequenzen nicht aufgrund ihrer wahren Verwandt-

schaftsbeziehungen gruppiert, sondern es wird stattdessen eine artifizielle Position im Baum unabhängig eines zugrunde liegenden phylogenetischen Signals eingenommen. Diese nicht-phylogenetischen Signale werden durch unerkannte oder falsch abgeleitete multiple Substitutionen generiert (Philippe *et al.*, 2011).

Stiller und Hall (1999) fanden beispielsweise Anzeichen dafür, dass die basale Topologie im Stammbaum der Eukaryoten abgeleitet aus ribosomalen DNA (rDNA)-Sequenzen vollständig als ein Artefakt der Variation molekularer Evolutionsraten zwischen den eukaryotischen Taxa erklärt werden könnte. Es besteht also in der Praxis immer die Möglichkeit, dass ein rekonstruierter Stammbaum eventuell fehlerbehaftet sein könnte, vor allem die Positionierung tiefer Äste in einer Phylogenie ist unverlässlich (Kumar und Rzhetsky, 1996).

Durch LBA kann sogar die Positionierung einer zur Wurzelung eines gegebenen phylogenetischen Baumes gewählten Aussenseitergruppe (engl. *outgroup*) beeinträchtigt werden (Stiller und Hall, 1999). Taxa, welche erheblich schneller evolvieren als andere, werden durch eine über einen sehr langen Ast verbundene Außenseitergruppe angezogen. Sie werden dadurch fälschlicherweise an der Basis des Baumes positioniert (Philippe *et al.*, 2011; Philippe und Laurent, 1998).

Grund für die Manifestierung von LBA innerhalb einer Phylogenie kann sein, dass parallele Austausche häufiger sind als informative nicht parallele Austausche (Felsenstein, 1978). Typisch ist das Auftreten von zahlreichen einzigartigen Sequenzaustauschen für bestimmte Sequenzen an ansonsten konservierten Positionen. Hierdurch entstehen viele scheinbare synapomorphe Positionen, die vor allem in Datensätzen mit limitiertem phylogenetischen Signal zu den bekannten falschen basalen Gruppierungen führen (Stiller und Hall, 1999). Das Verfahren der maximalen Parsimonie ist anfälliger für dieses Phänomen als die anderen beschriebenen Verfahren, jedoch sehr zuverlässig, falls diese beschriebenen Parallelismen nur vereinzelt vorhanden sind (Felsenstein, 1978).

Die Verzerrung der Nukleinsäure- beziehungsweise der Aminosäurenzusammensetzung (engl. *compositional bias*) als weiterer systematischer Fehler kann sogar vollständig unterstützte Bäume auf der Basis des *Bootstrap*-Verfahrens generieren, welche dennoch falsch sind (Phillips, Delsuc und Penny, 2004). In diesen gruppieren sich Sequenzen mit einer ähnlichen Zusammensetzung unabhängig von ihrer evolutionären Geschichte. Für mitochondrielle Genome konnte gezeigt werden, dass die Rekodierung der Nukleotide in Purine und Pyrimidine das phylogenetische Signal gegenüber diesem Fehler erhöhen konnte (Phillips, Delsuc und Penny, 2004).

Eine Möglichkeit, das Auftreten von sehr langen Ästen in einer Phylogenie zu vermeiden, besteht darin, mehr Sequenzen hinzuzufügen, so dass diese Äste im besten Fall in kürzere aufgebrochen werden. Mit Beginn des Zeitalters der Phylogenomik entstand die Hoffnung, die zahlreichen unstimmigen Ergebnisse aus phylogenetischen Analysen einzelner oder weniger Gene durch die stetige Addition weiterer Sequenzen miteinander in Einklang bringen zu können. Diese konnte sich jedoch bis heute nicht erfüllen (Philippe *et al.*, 2011).

Folgen die Ereignisse der Artentstehung bezogen auf den gesamten evolutionären Zeitraum relativ kurz hintereinander, ist es schwer, die zeitliche Abfolge dieser Ereignisse korrekt zu rekonstruieren (Philippe *et al.*, 2011; Philippe, Chenuil und Adoutte, 1994). Unter Umständen reicht die Anzahl der betrachteten Sequenzzeichen nicht aus, um die Phylogenie in einem solchen Bereich vollständig aufzulösen (Saitou und Nei, 1986). Die Auflösung dieser kurz auf einander folgenden Ereignisse kann ebenfalls durch eine unvollständige Verteilung von Merkmalen auf die Arten (engl. *incomplete lineage sorting*) (Philippe *et al.*, 2011) verhindert werden.

Die heutzutage häufig auftretenden Unstimmigkeiten zwischen Stammbäumen unterschiedlicher Gensammlungen werden gern als Zeichen vorherrschenden lateralen Gentransfers oder anderen nicht baumartigen biologischen Prozessen gedeutet. Jedoch treten diese Diskrepanzen häufig auch dann auf, wenn genau bekannt ist, wie die entsprechenden Moleküle evolvieren. White *et al.* (2007) konnten anhand eines Datensatzes plastidärer Gene, für welchen lateraler Gentransfer und paraloge Sequenzen weitestgehend vernachlässigt werden konnten, zeigen, dass das phylogenetische Signal mit zunehmender Divergenz der Sequenzen abnimmt und sich bei tiefen Divergenzen ähnlich zu randomisierten Datensätzen verhält. Vor allem die Informationen über interne Äste gehen dabei verloren, während der Anteil nicht-phylogenetischen Signals größer wird. Bei der Verwendung zusammengefügter Alignments konnte die Vergrößerung der nicht-phylogenetischen Signale im Gegensatz zur Betrachtung einzelner Gene zum Teil verhindert werden. Das hierbei verbliebene Signal hatte in diesem Fall jedoch ebenfalls wenig mit der korrekten Phylogenie zu tun. Dieser Verlust des phylogenetischen Signals mit steigender Sequenzdivergenz kann zum Teil tiefe Divergenzen mit vermeintlich falschen jedoch hinreichend unterstützten Topologien (Martin *et al.*, 1998) erklären.

Zur Rekonstruktion eines phylogenomischen Stammbaums kann man alle Genfamilien betrachten, welche universell in allen untersuchten Organismen vorhanden sind (Lukjancenko, Wassenaar und Ussery, 2010). Diese sogenannte Stammmenge

an Genen (engl. *core genes*) kann dann beispielsweise in einer Supermatrix zusammengefügt und analysiert werden. Durch das strikte Kriterium der Universalität reduziert man die zu betrachtenden Sequenzen sehr stark. Durch zusätzlichen Ausschluss von Genen, die in ihrer Geschichte lateralem Transfer unterlagen, konnte in einem repräsentativen Fall zur Rekonstruktion eines Stammbaums aller Lebewesen auf diese Weise nur etwa ein Prozent der Protein-kodierenden Gene eines durchschnittlichen Organismus bei der phylogenetischen Rekonstruktion berücksichtigt werden (Ciccarelli *et al.*, 2006; Dagan und Martin, 2006). Diese Methode hat jedoch auch ohne die Einbeziehung lateraler Transferereignisse einen entscheidenden Nachteil: Je mehr Taxa in die Analyse einbezogen werden, desto mehr schrumpft die Stammmenge an Genen, die betrachtet werden können.

Ein Artikel von Lukjancenko, Wassenaar und Ussery (2010) verdeutlicht das Ausmaß dieses Effekts sehr anschaulich. In ihrer Vergleichsanalyse von 61 sequenzierten Genomen nur einer einzigen Art (*Escherichia coli*) ermittelten sie einen Anteil von lediglich 6 % an universellen Familien unter allen im Pangenom enthaltenen Genfamilien. Ein derart geringer Anteil an Genen kann daher nur mit Vorsicht für die Evolution aller betrachteten Gene beziehungsweise der betrachteten Organismen an sich herangezogen werden (Dagan und Martin, 2006).

Minimierung von Fehlerquellen

Wir sehen also, dass in jedem Schritt der phylogenetischen Rekonstruktion von Stammbäumen unabhängig voneinander systematische Verzerrungen beziehungsweise Fehler eingeführt werden können. Nachfolgende Schritte basieren jeweils auf den Ergebnissen der vorherigen Schritte. Die Stammbaumrekonstruktion hängt beispielsweise von einer verlässlichen Clusteranalyse und der Berechnung eines korrekten Sequenzalignments ab. Fehler in vorauslaufenden Schritten werden tendenziell noch verstärkt.

Es ist angebracht diese inhärenten Fehlerquellen bei der Analyse weitestgehend zu minimieren. Zusätzlich zu einer sinnvollen Auswahl an Sequenzen verschiedener Organismen müssen die Limitierungen der einzelnen Verfahren, angefangen bei der beschriebenen Clusteranalyse, immer bewusst sein. Für das Alignmentverfahren ist es vorteilhaft, eine Methode auszuwählen, welche neueste Entwicklungen und Forschungsergebnisse auf diesem Gebiet mit einbezieht. Da die Beurteilung der Güte eines Alignments in der Praxis nur schwer eingeschätzt werden kann, sind Verfahren hilfreich, welche die Abschätzung der Verlässlichkeit der ali-

gnierten Positionen zulassen, wie dies zum Beispiel für die *Heads-or-tails*-Methode oder das *Guidance*-Verfahren (Landan und Graur, 2007; Penn *et al.*, 2010) der Fall ist. Die ermittelten Ergebnisse aus den Analysen auf Basis dieser Alignments müssen dann optimalerweise jeweils in Zusammenhang mit diesen möglichen Artefakten im Alignment in Verbindung gebracht werden (Abschnitt 5.3).

Zur phylogenetischen oder phylogenomischen Rekonstruktion von Genstammbäumen sollte zunächst ein für die Daten passendes Evolutionsmodell ausgewählt werden (Abschnitt 5.3). Zudem ist für die Auflösung tiefster Divergenzen in Stammbäumen eine möglichst korrekte Behandlung der Insertions- und Deletionsereignisse und die anschließende Anwendung eines phylogenetischen Rekonstruktionsverfahrens, welches diese Regionen überhaupt berücksichtigt, wünschenswert. Zu Berücksichtigen ist auch die Frage, inwieweit lateraler Gentransfer die zu beobachtenden Daten beeinflusst haben könnte, um keine falschen Rückschlüsse zu ziehen. Ist ein phylogenetischer Baum überhaupt ein adäquates Mittel, die vorhandenen Daten zu beschreiben oder sollten andere Methoden der Darstellung in Betracht gezogen werden? In einigen Fällen kann es sehr hilfreich sein, eine gegebene Phylogenie mit einem phylogenetischen Netzwerk derselben Sequenzen zu vergleichen, um diesen Einfluss abzuschätzen (Abschnitt 5.3, Abschnitt 6.3). Unter Umständen können alternative Lösungsansätze lange offene Fragen beantworten, welche mit Hilfe der Standardmethoden immer wieder zu widersprüchlichen Ergebnissen geführt haben (Abschnitt 5.1, Abschnitt 6.1).

Joseph Felsenstein formulierte 1978 einen Satz, der auch vier Jahrzehnte später noch seine Gültigkeit hat, und beschreibt, dass man einer phylogenetischen Methode erst dann vertrauen sollte, wenn ihre einwandfreie Funktion restlos bewiesen ist: „If phylogenetic inference is to be a science, we must consider its methods guilty until proven innocent“(Felsenstein, 1978).

4 Zielsetzung

Die Aufklärung biologischer Zusammenhänge erfordert oftmals die Rekonstruktion sehr weit zurückliegender Evolutionsereignisse. Beispiele hierfür sind die Frage nach der frühesten Aufteilung im Stammbaum der Prokaryoten und damit im Stammbaum aller lebenden Organismen. Auch die Frage, welches rezente Cyanobakterium die höchste genomische Ähnlichkeit zum gemeinsamen Vorfahren aller Plastiden hat betrifft das weit zurückliegende Ereignis der primären Endosymbiose. Die Rekonstruktion dieser frühen Ereignisse beinhaltet standardmäßig die Berechnung tiefer phylogenomischer Stammbäume und deren Interpretation. Je tiefer die betrachteten Divergenzen in einer Phylogenie liegen, desto anfälliger ist deren Ableitung jedoch für Rekonstruktionsfehler.

Die in dieser Arbeit angewendeten Methoden sollten anhand der genannten Beispiele nach alternativen Lösungswegen suchen, welche bekannte Fehlerquellen möglichst vermeiden sollten. Neben der Aufklärung der evolutionären Ereignisse war es insbesondere das Ziel, die Betrachtung möglicher Fehlerquellen mit in die Auswertung einzubeziehen, welche im Zusammenhang mit Unsicherheiten im Sequenzalignment als grundlegendem Schritt phylogenetischer und phylogenomischer Analysen stehen. Die Ergebnisse sollten unter diesen Aspekten kritisch betrachtet werden. Des Weiteren sollte untersucht werden, in welchem Ausmaß problematisch zu alignierende Alignments aus sich heraus die Ableitung unstimmiger Genbäume im Vergleich mit einem Stammbaum der zugrundeliegenden Spezies negativ beeinflussen, also zu fehlerhaft rekonstruierten Bäumen führen können.

5 Publikationen

5.1 Genome networks root the tree of life between prokaryotic domains

Tal Dagan[1,⋆], Mayo Roettger[1,⋆], David Bryant[2], and William Martin[1]

[1]Institut für Botanik III, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

[2]Department of Mathematics, University of Auckland, Auckland, New Zealand

[⋆]Der Beitrag beider Autoren ist gleichwertig.

Der vorgestellte Artikel wurde in der Fachzeitschrift *Genome Biology and Evolution* veröffentlicht (Dagan *et al*., 2010).

Beitrag von Mayo Röttger, gleichberechtigter Erstautor:

Versuchsplanung:	50 %
Datenanalyse:	80 %
Verfassen des Manuskripts:	30 %

GENOME BIOLOGY AND EVOLUTION

SMBE

GBE

Genome Networks Root the Tree of Life between Prokaryotic Domains

Tal Dagan*,†,[1], Mayo Roettger†,[1], David Bryant[2], and William Martin[1]

[1]Institute of Botany III, Heinrich-Heine University of Düsseldorf, Düsseldorf, Germany

[2]Department of Mathematics, University of Auckland, Auckland, New Zealand

*Corresponding author: E-mail: tal.dagan@uni-duesseldorf.de.

†These authors contributed equally to this work.

Accepted: 12 May 2010

Abstract

Eukaryotes arose from prokaryotes, hence the root in the tree of life resides among the prokaryotic domains. The position of the root is still debated, although pinpointing it would aid our understanding of the early evolution of life. Because prokaryote evolution was long viewed as a tree-like process of lineage bifurcations, efforts to identify the most ancient microbial lineage split have traditionally focused on positioning a root on a phylogenetic tree constructed from one or several genes. Such studies have delivered widely conflicting results on the position of the root, this being mainly due to methodological problems inherent to deep gene phylogeny and the workings of lateral gene transfer among prokaryotes over evolutionary time. Here, we report the position of the root determined with whole genome data using network-based procedures that take into account both gene presence or absence and the level of sequence similarity among all individual gene families that are shared across genomes. On the basis of 562,321 protein-coding gene families distributed across 191 genomes, we find that the deepest divide in the prokaryotic world is interdomain, that is, separating the archaebacteria from the eubacteria. This result resonates with some older views but conflicts with the results of most studies over the last decade that have addressed the issue. In particular, several studies have suggested that the molecular distinctness of archaebacteria is not evidence for their antiquity relative to eubacteria but instead stems from some kind of inherently elevated rate of archaebacterial sequence change. Here, we specifically test for such a rate elevation across all prokaryotic lineages through the analysis of all possible quartets among eight genes duplicated in all prokaryotes, hence the last common ancestor thereof. The results show that neither the archaebacteria as a group nor the eubacteria as a group harbor evidence for elevated evolutionary rates in the sampled genes, either in the recent evolutionary past or in their common ancestor. The interdomain prokaryotic position of the root is thus not attributable to lineage-specific rate variation.

Key words: phylogenies, early evolution, tree of life, microbial genomics, lateral gene transfer.

Introduction

Geochemical and isotopic data indicates that life on earth was already flourishing by the time that the oldest known sedimentary rocks had formed some 3.5 Ga (Ueno et al. 2006) and that by about 3.2 Ga prokaryotic communities in anaerobic marine environments looked very much like today's (Nisbet 2000; Rasmussen 2000; Shen et al. 2001; Brasier et al. 2006; Grassineau et al. 2006). Microfossil data reflect a more or less continuous record of abundant prokaryotic communities from ~3.5 Ga onward, with eukaryotes appearing later. The presence of diversified and unequivocally eukaryotic cells is documented in sediments ~1.5 Ga of age (Javaux et al. 2001; Knoll et al. 2006), followed by eukaryotic algae at ~1.2 Ga (Butterfield 2000). Biomarker evidence once suggested the possible presence of eukaryotes by 2.7 Ga, but the biomarkers were subsequently shown by virtue of their isotope fingerprint not to have arisen within the rocks in which they occur (Fischer 2008; Rasmussen et al. 2008). Accordingly, eukaryotes appear about 2 billion years later in the geological record than do prokaryotes, consistent with the results of recent molecular and genomic investigations indicating that eukaryotes, which ancestrally possess mitochondria, arose from prokaryotes, lineages to which both mitochondria and their host

can be traced (Rivera and Lake 2004; Embley and Martin 2006; Pisani et al. 2007; Cox et al. 2008; Koonin 2009). Thus, the first 2 billion years of life on earth, in particular, the very first phases of life's history, are about prokaryote evolution only. Hence the position of the root in the "tree of life" concerns the deepest divide among the prokaryote groups.

Early efforts to locate the root in the tree of life focused on phylogenies of individual genes (Gogarten et al. 1989; Iwabe et al. 1989; Brown and Doolittle 1995). But with the recognition of lateral gene transfer (LGT) as a widespread and altogether normal mechanism of natural variation affecting prokaryote genome evolution (Doolittle 1999; McInerney and Pisani 2007), concerns became increasingly severe, and well founded, that any individual gene could serve as a reliable proxy for the evolution of a whole genome all the way back to the earliest divergence events in life's history.

More recently, indels present in seven anciently conserved proteins (IF2, EF-G, Hsp70, HisA, S12, GyrA, PyrD) have been used to infer the position of the root (Lake et al. 2009). This approach first excluded the root from the archaebacteria (Skophammer et al. 2006), then from the Gram-negative eubacteria (Lake et al. 2007) and finally placed it within the eubacteria, on a branch separating the firmicutes and the archaebacteria from all else (Lake et al. 2008). These studies were, however, criticized on the basis that the alignments were problematic (Di Giulio 2007). Furthermore there is the issue that seemingly robust indels can in fact arise independently at the same spots of a protein alignment (and structure) during evolution (Bapteste and Philippe 2002). In addition, the LGT caveat holds for the indel data as well, that is, it is highly questionable whether the evolutionary patterns preserved in the indels of any one gene are indicative for the evolution of the entire genome. Indeed, it is presently difficult at best to muster evidence that any gene has remained immune to lateral transfer over the fullness of geological time (Bapteste et al. 2009). Moreover, the approach to phylogeny using indels, as extensively applied by Gupta and colleagues over the years (Gupta 1998; Gupta and Lorenzini 2007), has the drawback that rather than looking at all the indels, which would contain a large amount of conflicting data, one only looks at a few specifically chosen indels, giving the impression that indel data lack substantial conflict.

Another recent approach to inferring the position of the root in the tree of life entails the logical-parsimonious analysis of characters (Cavalier-Smith 2006b). However, that approach entails a dismissal of molecular data from genomes as inapplicable to the study of microbial evolution because it allows lineage-specific and gene-specific variations of evolutionary rate to be assumed without penalty by invoking "quantum evolution" wherever convenient to account for any observed pattern of sequence similarity or lack thereof (Cavalier-Smith 2010b). As such, the method is independent of tests with evidence founded in gene sequence similarity. It nonetheless places the root within the Chloroflexi, anoxygenic photosynthetic eubacteria (Cavalier-Smith 2010a), and prescribes an origin of the archaebacteria (and eukaryotes) from actinobacteria only 850 Ma (Cavalier-Smith 2006a). That suggestion is distinctly at odds with geochemical evidence for biological methane production >3 Ga (Canfield 2006; Ueno et al. 2006), with biomarker evidence for archaebacteria in 2.7 Ga deposits (Ventura et al. 2007) and is difficult to reconcile with the observation that many archaebacteria inhabit hydrothermal niches that have existed for as long as there has been water on earth (Sleep et al. 2004). It is furthermore at odds with unequivocal microfossil evidence for the existence >850 Ma of eukaryotes (Butterfield 2000; Javaux et al. 2001), which in the neomuran theory are viewed as descendants of the same actinobacterial group as archaebacteria.

Genome-wide data deliver yet other distinctly differing results with respect to the position of the root. Wong et al. (2007), for example, used a combination of data types in an analysis that placed the root close to *Methanopyrus* within the archaebacteria. That rooting is consistent with isotope evidence for the antiquity of methanogenesis (Ueno et al. 2006). In other work, Zhaxybayeva et al. (2005) analyzed 12 anciently duplicated gene pairs and concluded that the root probably lies between the archaebacteria and the eubacteria but pointed to the caveat that 12 genes might not speak for the whole genome because of LGT and furthermore pointed out a lack of strong phylogenetic signal in their data. Boussau et al. (2008) investigated rRNA phylogeny and about 50 proteins also concluded that the root probably lies between archaebacteria and eubacteria. Indeed, various authors embrace the view that the root lies between archeabacteria and eubacteria because of the few molecular characters that these groups share in common in their genome comparisons (Dagan and Martin 2007; McInerney et al. 2008; Battistuzzi and Hedges 2009; Koonin 2009) but without providing specific molecular analyses to support that view.

Specific attempts to root the tree of life through data analyses deliver conflicting results, although most commonly a eubacterial root (Gogarten et al. 1989; Lake et al. 2009). Particularly problematic with any rooting of the tree of life within the eubacteria, however, is that the archaebacteria—which 1) generally share very few genes with eubacteria (Snel et al. 1999; Graham et al. 2000), 2) have different plasma membrane and cell wall chemistries than eubacteria (Martin and König 1996; Claus et al. 2005; Engelhardt 2007), 3) have different machineries of DNA maintenance than eubacteria (Chong et al. 2000; Frols et al. 2009), 4) employ many different cofactors than eubacteria (Dimarco et al. 1990; Deppenmeier 2002; Fujihashi et al. 2007), and 5) have different core promotor and RNA polymerase

structures than eubacteria (Bell and Jackson 2001)—assume the status of a derived group of eubacteria in such schemes. Importantly, all current eubacterial root views (Lake et al. 2008; Cavalier-Smith 2010b) invoke the hitherto untested corollary assumption that there is some form of systematic acceleration in the evolutionary rate of sequence change within the archaebacterial lineage. Such studies are furthermore based on a few specifically chosen characters, not whole genome data.

Genomes sequences should contain more evidence addressing the deepest divide among prokaryotes than just a few genes do. The root inferred from whole genomes should correspond to the bipartition separating those genomes that share the fewest genes in common and the least sequence similarity. That root should, in turn, correspond to the most ancient, in terms of geological time, split in the prokaryotic world, barring the existence of lineage-specific rate fluctuations across that divide, a hefty caveat. Here, we pinpoint the most ancient prokaryote genome divergence on the basis of whole genome data. By analyzing gene distribution patterns, we reconstruct a phylogenetic network of 191 prokaryotes. Using the midpoint rooting approach (Farris 1972), we then identify the root position within the network. Furthermore, we show through quartet analysis of the eight ancient paralogous genes that arose by duplication in the prokaryote common ancestor that the position of the root so identified cannot be attributed to lineage-specific increases in rates of sequence change.

Materials and Methods

Orthologous Protein Families Completely sequenced prokaryotic genomes were downloaded from the National Center for Biotechnology Information (NCBI) Website (http://www.ncbi.nlm.nih.gov/; genomes available at August 2005). For each species, only the strain with the largest number of genes was used. Of 191 genomes (562,321 proteins) in the data, 22 are archaebacterial and 169 are eubacterial. All proteins in the 191 genomes were clustered by similarity into gene families using reciprocal best Blast hit (BBH) approach (Tatusov et al. 1997). Each protein was Blasted against each of the genomes. Pairs of proteins that resulted as reciprocal BBHs of E value $< 10^{-10}$ were aligned using ClustalW (Thompson et al. 1994) to obtain amino acid identities. Protein pairs with $\geq$30% amino acid identity where clustered into protein families of $\geq$2 members using the Markov cluster algorithm (MCL; Enright et al. 2002) setting the inflation parameter, I, to 2.0. For the comparison of gene distribution patterns over different protein similarity thresholds, six additional sets of protein families were clustered using ascending threshold (T_i, where i = {35,40,45,50,55,60}) for the percent amino acid identity between protein pairs that are included in the analysis. Protein families reconstructed by the MCL algorithm include both orthologous and paralogous proteins. Because, in this study, we are interested in orthologous proteins only, we sorted out the paralogous genes from the protein families. To distinguish between orthologs and paralogs, we used the number of reciprocal BBHs for each gene within a family. In the case of multiple genes for a genome in a certain protein family, orthologs are expected to have more reciprocal BBHs in other genomes than paralogs. Thus, for each genome, only the protein with the maximum number of reciprocal BBHs is considered.

Splits Network Protein families from each protein similarity threshold were compared with the protein families reconstructed under a 5% higher threshold. Proteins that are included within one family at a certain threshold may be clustered into one or more families at the higher threshold. The first case indicates a conservation of the family and the latter indicates one split or more. Thus, for each of the families in the higher threshold (those that comprise proteins clustered into a single family at the lower threshold), a new split is recorded in a binary pattern that includes 191 digits; if the protein family includes a protein from genome i then digit x_i in its corresponding pattern is "1," otherwise it is "0." Species that are not represented in the protein family are coded as "?". All splits for a certain threshold were then summarized by a splits network using SplitsTree (Huson and Bryant 2006).

Midpoint Rooting in Splits Network The root within the splits network was located by adapting the midpoint rooting approach in phylogenies (Farris 1972). This method assumes that all lineages evolve at roughly similar rates. In a phylogeny, the root is located half way along the path connecting the pair of taxa that are furthest apart in the tree (fig. 1*a*). Here, the distance between two taxa in the tree is measured according to "phenetic distance," the length of the path (i.e., the sum of split weights) from one taxon to the other in the tree.

In a split network, there can be multiple paths between any two nodes, and the phenetic distance between two nodes in a split network is therefore defined as the length of the "shortest" path connecting the nodes. As well, there can be multiple shortest paths between two nodes, giving multiple possible midpoint locations (fig. 1*b*).

To locate the root of the split network, a pair of taxa at maximum phenetic distance is identified. Ties can be broken arbitrarily: any pair with the maximum distance will give the same root location. Once a pair is selected, the set of path midpoints half way between the two taxa is obtained. An arbitrary reference taxon v is selected, and the splits in the network are numbered 1, 2, . . . , m. The location of each midpoint node x is then encoded as a vector $(x_1, x_2, \ldots, x_m)$ of length m where $x_i = 1$ if the shortest paths from v to x traverse an edge labeled by split i and $x_i = 0$ if they don't. In a split network, all the shortest paths between any two

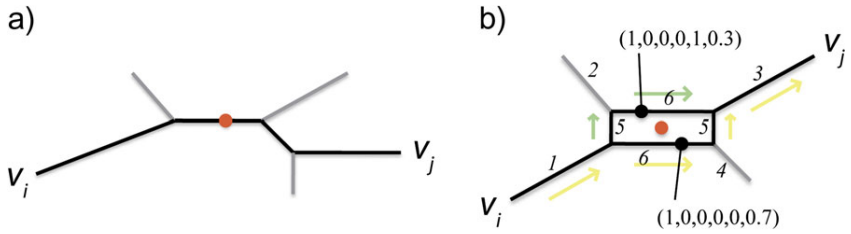


FIG. 1.—Midpoint rooting trees and networks. (*a*) The pathway from node v_i to v_j and the midpoint (red circle) in a phylogenetic tree is shown. (*b*) An illustration of the procedure used to root a split network. The two most distant taxa are v_i and v_j. There are two shortest paths between these two taxa (colored arrows) and two midpoints. The numbering of the splits is indicated, noting that the two central splits are associated with two edges each. The vector encoding is made with reference to taxon v_i. The encoding for the root is (1,0,0,0.5,0.5), which corresponds to the center of the central box (red circle).

nodes will cross over edges labeled by the same set of splits (Dress and Huson 2004). This encoding is extended to locations along edges or within boxes by allowing the components of the vector to take on fractional values between 0 and 1. Let $(z_1, z_2, \ldots, z_m)$ be the average of the midpoint location vectors; this is the location vector for the root. To determine the position of the root in the network, a path is traced starting from *v* and using edges labeled by splits *i* for which $z_i = 1$ (and never two edges with the same labels). The fraction components of the location vector then determine the position of the root along the edge or within a box.

It can be shown that in any planar drawing of the split network, the position of the root in the plane will be exactly the average of the positions of the midpoint. Also, when the split network is actually a tree, this network root will be exactly the midpoint root.

The robustness of the midpoint network root was tested using a type of jackknife resampling approach. By this approach, the most distant pair of taxa is excluded from the splits network, and the midpoint root is recalculated. This procedure was repeated until the root was no longer found between archaebacteria and eubacteria. We note that if a large number of pairs need to be removed to modify the position of the root then that position will also be stable if random taxa are removed according to a statistical jacknife procedure.

Test of the Global Clock Assumption Ancient paralogous genes were identified by their four-letter synonym within NCBI's genome annotations (ptt files). Genomes for which proteins were not found using the four-letter synonym were searched by reciprocal BBH procedure using an already identified protein from the same lineage (see below) as a query and the genome in question as subject. The annotation of proteins identified this way was double-checked manually. The taxonomic classification of the 191 species is done by NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy). For species within Firmicutes or Proteobacteria phyla, the lineage is defined as the taxonomic class otherwise it is defined as the taxonomic phylum.

Quartets of ancient paralogs (fig. 2*a*) were assembled from the sequences of two ancient paralogs from two different lineages for all possible species pairs. Sequence alignments were reconstructed using ClustalW (Thompson et al. 1994). Sequence alignment reliability was tested using the HoT procedure (Landan and Graur 2007), and only alignments with a sum-of-pairs score >80% were included in the analysis. Phylogenetic trees reconstructed from quartets of ancient paralogs may result in three possible topologies (fig. 2*b*). The most likely tree topology for each quartet was tested with the SH test (Shimodaira and Hasegawa 1999) using ProML of PHYLIP (Felsenstein 1996). Only quartets of vertical topology (t_{vert}) were considered for further analysis.

Different models of evolutionary rate variation along the branches were tested using the PAML package (Yang 2007). Each quartet was first tested for global molecular clock model (r_{global}), assuming equal rates on all branches (fig. 2*c*), using the null hypothesis H_0: all branches evolve with rate r_1. This model has three parameters corresponding to the $n - 1$ interior nodes in a rooted tree, whereas the alternative hypothesis H_1 assumes different rates for all five branches in an unrooted tree and therefore has five parameters for a tree of four taxa (Yoder and Yang 2000). The maximum log-likelihood values under both models (l_0 and l_1, respectively) are estimated with CodeML, and twice the log-likelihood difference, $2\Delta l = 2(l_1 - l_0)$ was compared with a χ^2 distribution with degrees of freedom (df) = 2 to test whether the global clock hypothesis is rejected (Yang 1998). Quartets for which the global clock hypothesis was rejected were subsequently tested for a lineage-specific rate ($r_{lineage}$) assuming different rates between the two lineages and equal rates between each paralogs pair (fig. 2*c*). The null hypothesis in this case is H_0: branches α_1, β_1 evolve with rate r_1 and branches α_2, β_2 evolve with rate r_2. The alternative hypothesis H_1 assumes the free-rate model again with its five parameters. Because the lineage-specific rate model has two free parameters less than the free-rate model, we analogously compare $2\Delta l = 2(l_1 - l_0)$ with a χ^2 distribution with df = 2 to test whether the null hypothesis is rejected.

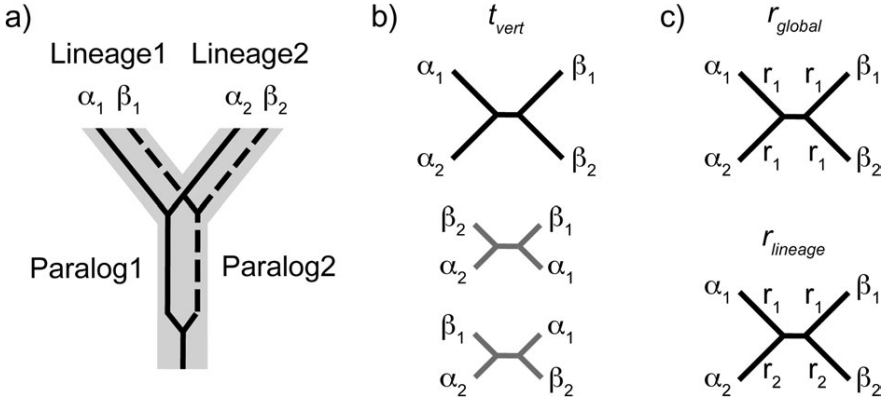


Fig. 2.—Ancient paralogs quartet analysis. (*a*) Ancient paralogs are defined as paralogous proteins that were duplicated in the common ancestor of archaebacteria and eubacteria. (*b*) For a phylogenetic tree of four OTUs (operational taxonomic units), there are three possible topologies. No LGT among the major taxa results in topology t_{vert} (in black), whereas evolution by LGT may result in any of the other two topologies (in gray). (*c*) Here, we tested two different rate models for the t_{vert} topology: in the r_{global} model all OTUs evolve in the same rates. In the $r_{lineage}$ model, OTUs from the same lineage evolve in the same rate, which differs between the lineages.

Results and Discussion

Splits Networks for Prokaryotic Genomes We clustered the 562,321 protein-coding sequences that occur among 191 completely sequenced prokaryotic genomes from 15 higher level taxa with standard procedures (Enright et al. 2002) into groups based upon sequence similarity threshold. The clustering threshold corresponds to a value of amino acid identity, T_{30} designating the 30% threshold, for example, indicating that each protein in the cluster at T_{30} must share at least 30% amino acid identity with one other member (not with all other members) in the cluster. Depending upon the threshold set for the clustering procedure, these proteins fall into comparatively few large and inclusive families of distantly related sequences, or many smaller families whose members share high sequence identity. For example, clustering using T_{30} results in 57,743 families, 103 of them are nearly universal, including ≥90% of the species and 39,781 families include between 2 and 4 species.

Across different clustering thresholds T_i of increasing stringency (in 5% increments, e.g., T_{30}, T_{35}, T_{40}, etc.), a given family will tend to break apart into two or more separate families, each containing a smaller number of more highly conserved sequences at higher values of T_i. Depending upon the distribution of sequence similarities within a given family, an individual increase in clustering stringency, ΔT_i, may or may not introduce such a split within the protein family. Each new split within the family, termed here a protein family split, corresponds to a split among the strains (genomes) in which the family is present, termed here a genome split. The set of all genome splits can be readily converted into networks using NeighborNet (Bryant and Moulton 2004) in SplitsTree (Huson and Bryant 2006), which constructs phylogenetic networks based on the Neighbor-Joining algorithm (Saitou and Nei 1987). In the resulting networks, splits separating the genome set reflect overall sequence similarity between members of all protein families shared across the corresponding genomes, regardless of whether that similarity stems from vertical descent, differential loss, or LGT.

At the amino acid identity threshold of 25% (T_{25}), the 562,321 proteins fall into 53,429 families of ≥2 proteins. Of those, only 3,832 (7% of the total) of the families have members occurring in both archaebacteria and eubacteria. The fraction of protein families with this broad distribution decreases with the increase of protein similarity threshold, down to 172 (0.2%) in T_{60}. The fraction of archaebacterial-specific proteins remains almost constant (10–11%) across values of T_i, whereas the proportion of eubacterial-specific proteins increases from 83% to 90%. The proportion of group-specific protein families increases with the protein similarity threshold in most groups (e.g., Actinobacteria and α-Proteobacteria), whereas in Cyanobacteria this proportion is almost constant (supplementary table S1, Supplementary Material online). Hence, reconstruction of protein families using ascending amino acid identity threshold generally yields more exclusive protein families of increasingly narrow taxonomic range. Moreover, when increasing the protein similarity threshold, inclusive protein families (e.g., proteobacterial specific) split into more exclusive protein families (e.g., α- and β-Proteobacteria).

The set of all protein family splits was then extracted by comparison of families clustered at incrementally increased thresholds. To illustrate, at T_{25} only six protein families are present in all 191 species in the data set (fig. 3). Three of them—translation elongation factor G (fig. 3*a*), ribosomal protein L1 (fig. 3*b*), and ribosomal protein L5 (fig. 3*c*)—split

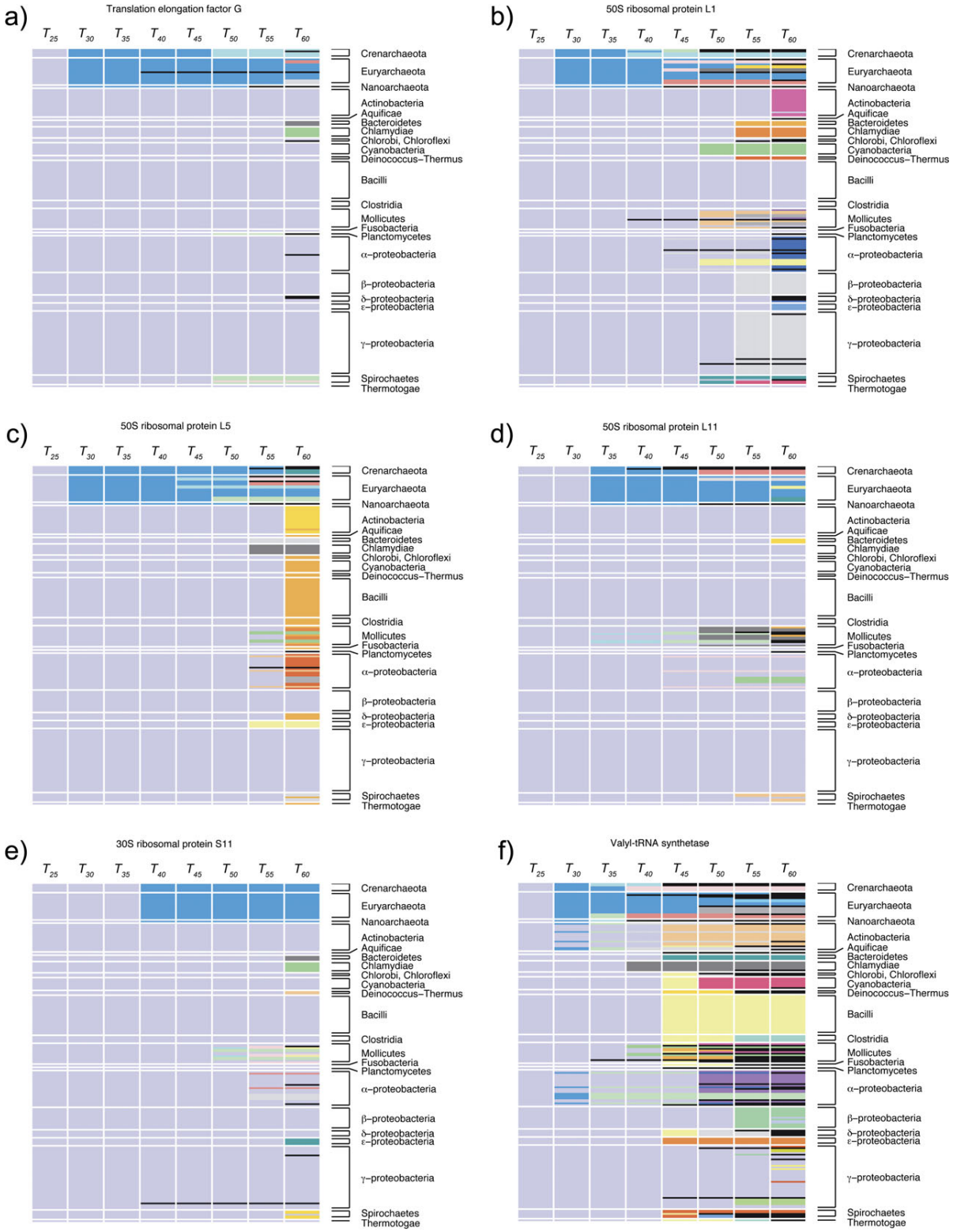


Fig. 3.—Protein family splits over ascending protein similarity thresholds for six protein families that are universal at T_{25}: (*a*) Translation elongation factor G, (*b*) 50S ribosomal protein L1, (*c*) 50S ribosomal protein L5, (*d*) 50S ribosomal protein L11, (*e*) 30S ribosomal protein S11, (*f*) valyl-tRNA synthetase. The splits are shown as colored boxes within columns. Currently recognized taxonomic groups are indicated in rows for comparison. For example, 50S ribosomal protein L5 (*c*) is universal at T_{25}, whereas in T_{30} the protein family splits into an archaebacteria-specific family (blue) and a eubacteria-specific family (light purple).

Genome Biology and Evolution

SMBE

into two protein families: one archaebacterial specific and one eubacterial specific. Only two of these families are still universal at T_{30}: ribosomal proteins L11 (fig. 3*d*) and S11 (fig. 3*e*). The last family, valyl-tRNA synthetase (ValRS), splits at T_{30} into one eubacterial-specific family and one including all archaebacteria, five Actinobacteria, and seven α-Proteobacteria (fig. 3*f*). At T_{35}, the latter ValRS family splits into three families, two of them containing archaebacteria only and one including the Thermoplasmatales (Euryarchaeota), Actinobacteria, and α-Proteobacteria. At T_{40}, the latter family splits into three families specific to Thermoplasmata, Actinobacteria, and α-Proteobacteria, respectively. These splits are the result of lateral transfer of ValRS genes from archaeabacteria to α-Proteobacteia and Actinobacteria (Raoult et al. 2003), followed by vertical descent within these groups.

Prokaryotic Genome Clusters Comparison of networks obtained from different protein similarity thresholds shows that ancient genome splits contain more information about divergence of the major taxa than recent ones (fig. 4). This is because using higher protein similarity thresholds results in an increased proportion of taxon-specific families (supplementary table S1, Supplementary Material online), a shift of the split information to the tips of the network, and as a result, a collapse of the network into a star-like topology (supplementary fig. S1, Supplementary Material online). Overall, the protein family split networks tend to recover traditionally recognized prokaryotic groups at higher taxonomic levels (fig. 5). Splits of protein families in the lower thresholds, for example, the $T_{25} \rightarrow T_{30}$ splits and $T_{35} \rightarrow T_{40}$ splits, contain enough information to recover the divergence of the major prokaryotic groups, so that most of them are "monophyletic" in the sense of there being a split in the data that unites them to the exclusion of all other taxa irrespective of conflicting splits using figure 5. This is a somewhat liberal use of the word monophyletic in this context because it focuses on the criterion "is there any signal uniting them" as opposed to asking "does any signal divide them." A network is a composite of multiple potentially conflicting signals, and the presence of a split separating out a clade suggests (in an unrooted sense) the presence of at least some phylogenetic evidence in favor of the clade being monophyletic for at least part of the genome. It is notable that only three higher groups examined here failed that monophyly criterion at all thresholds: the proteobacteria, the euryarchaeotes, and the clostridia (fig. 6). This is worth a brief consideration.

In general, the lack of monophyly for groups in the present analysis is most easily attributed to patchy patterns of gene sharing across groups, for example, as afforded by LGT during evolution. That the proteobacteria are not monophyletic in our analyses is largely attributable to their frequency in the sample size and their general tendency to harbor large and diverse genomes with abundant LGT (Lang and Beatty 2007; Dagan et al. 2008). More curious is the lack of monophyly for the clostridia, which contains many acetogens (Pierce et al. 2008; Ljungdahl 2009) and the euryarchaeotes, where the methanogens reside (Thauer et al. 2008). Acetogens and methanogens are strict anaerobes and inhabit environments that have existed since there was first life on earth (Martin et al. 2008), they both gain their energy from the reduction of CO_2 with H_2, they both harbor forms that can generate their chemiosmotic ion gradients without the participation of cytochromes (Müller 2003) or quinones (Thauer et al. 2008; Biegel et al. 2009). The lack of monophyly might relate to the large amounts of gene exchange across higher taxa involving these groups, for example, as in the hundreds of clostridial genes found in Thermotogales (Zhaxybayeva et al. 2009), or the dozens (Chistoserdova et al. 1998) to hundreds (Deppenmeier et al. 2002) to thousands of genes (Ng et al. 2000) that have been exchanged between some euryarchaeotes and eubacteria. Another possible interpretation is that if LGT is as prevalent in the environment and over geological time as some are claiming (Doolittle and Bapteste 2007), then the oldest prokaryotic groups will have had the greatest opportunity to exchange genes with other groups hence, eroding their monophyly be the measure of whole genome comparison used here. In that sense, and with the corresponding caveats, the lack of monophyly for the clostridia and euryarchaeotes could reflect their antiquity relative to the other groups sampled here.

In the three most ancient split networks (fig. 4), archaebacteria are monophyletic but within this kingdom the euryarchaeotes are paraphyletic, consistent with the findings of other recent studies (Fukami-Kobayashi et al. 2007; Cox et al. 2008; Puigbò et al. 2009). Only three species out of the 191 genomes do not branch with their traditionally assigned taxonomic group within the splits networks (for details, see supplementary table S2, Supplementary Material online).

The Root of Prokaryotes The concept of rooting is familiar in the realm of phylogenetic trees but has so far not been developed in the context of phylogenetic networks. The simplest form of rooting entails finding the two most distance species and placing the root on their midpoint, but it also entails a global rate constancy assumption (Farris 1972). Midpoint rooting for a network must, however, take into account multiple paths between pairs of taxa. Here, the midpoints are calculated for all equally shortest paths between the two most distant species and then all midpoints are "averaged" into a new root location within the network (see Materials and Methods). The two most distant species in the $T_{25} \rightarrow T_{30}$ network are *Thermoplasma acidophilum* (Euryarchaeota) and *Mycoplasma pneumoniae* (Tenericutes). Averaging the midpoint among all shortest paths

Genome Biology and Evolution

SMBE

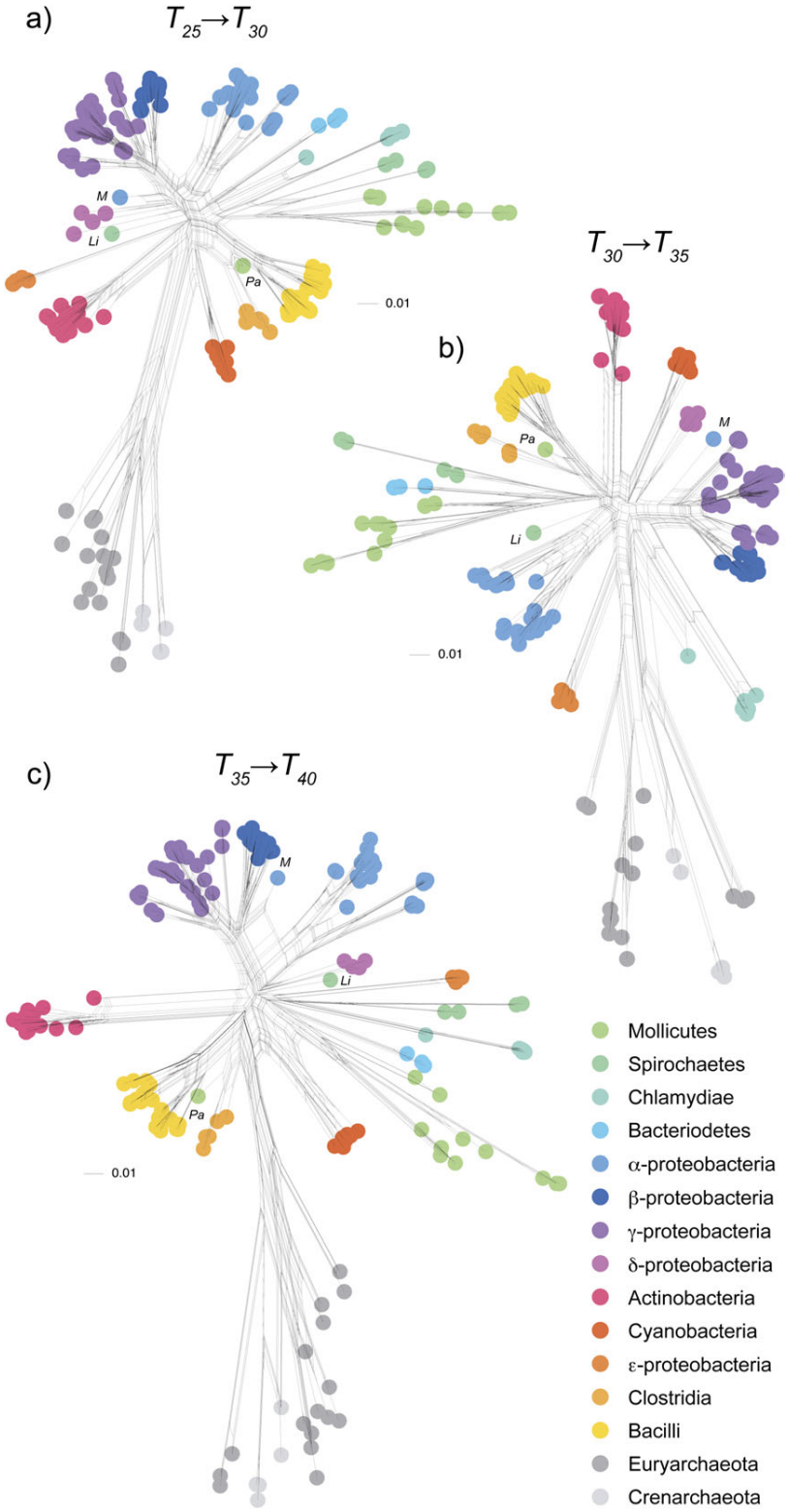


Fig. 4.—Protein family splits networks for the lowest three protein similarity cutoffs. Networks for higher protein similarity cutoffs are presented in supplementary figure S1 (Supplementary Material online).

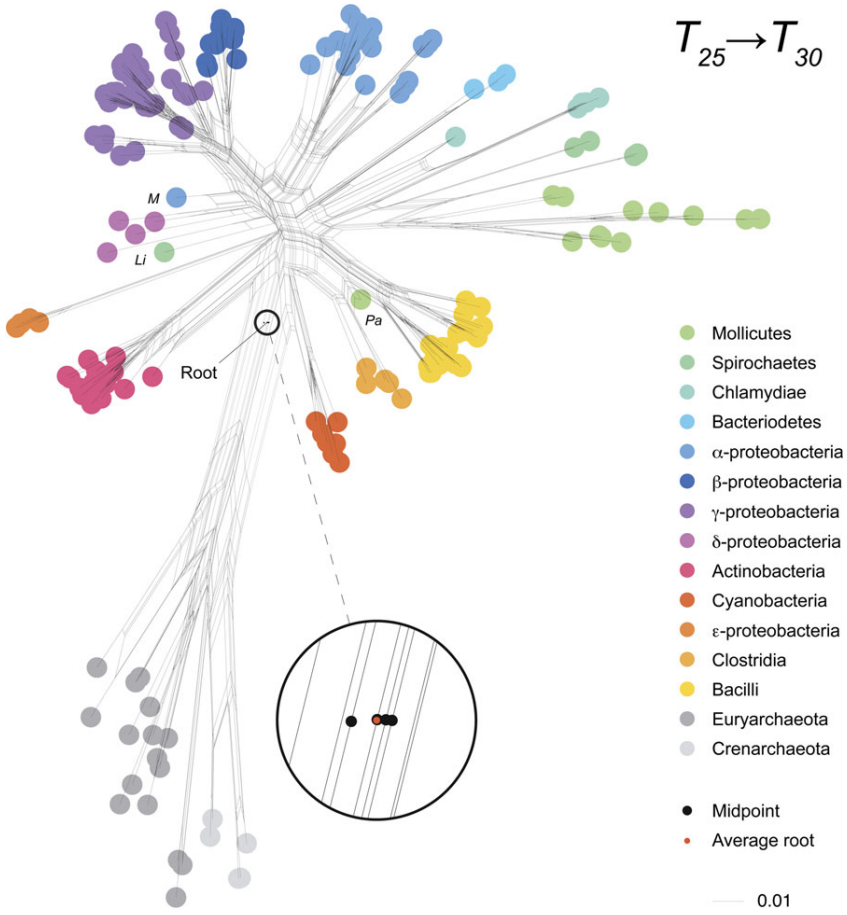


Fig. 5.—Midpoint root location in the $T_{25} \rightarrow T_{30}$ protein family splits network.

results in a root location on the split between archaebacteria and eubacteria (fig. 5).

In order to test the robustness of the root placement between archaebacteria and eubacteria, we applied a jackknife resampling approach to our network rooting procedure. In this approach, the rooting procedure is iterated, whereby in each iteration the most distant species from the previous iteration are excluded from the network until the result location of the root changes. Here, we repeated the rooting procedure until the root was no longer located between archaebacteria and eubacteria. The robustness of the root location is thus dependent on the number of iterations. The original placement of the root is between *T. acidophilum* and *M. pneumoniae*. After excluding those two species from the network, we find that the root is placed between *Sulfolobus acidocaldarius DSM 639* and *Mycoplasma genitalium*. Excluding the most distant pair in each step results in smaller distances as the iterations proceed (supplementary table S3, Supplementary Material online). After applying the exclusion and rerooting procedure iteratively for 20 times, we still find the root on the split separating archaebacteria (*Methanosarcina acetivorans*) from eubacteria (*Mycobacterium bovis*). Further exclusion of *M. acetivorans* as a member of the euryarchaeota group results in a network devoid of archaebacteria, rerooting of which places the root on a split between Actinobacteria and the remaining eubacteria.

The split networks reconstructed for increasing T_i also show that the split found in the rooted network is also the most ancient split among prokaryotes because it is the strongest split at the lowest amino acid identity thresholds and weakens when higher thresholds (more closely related proteins only) are queried (fig. 4, supplementary fig. S1, Supplementary Material online). However, just as with rooting trees, this approach to rooting the network can be sensitive to rate variation because split weight can be affected by variation in the rate of sequence change among groups. Hence, it was important to test for lineage- or genome-specific rate variation, which we did for 191 genomes using eight ancient paralogs that were duplicated in the common ancestor of genomes sampled here.

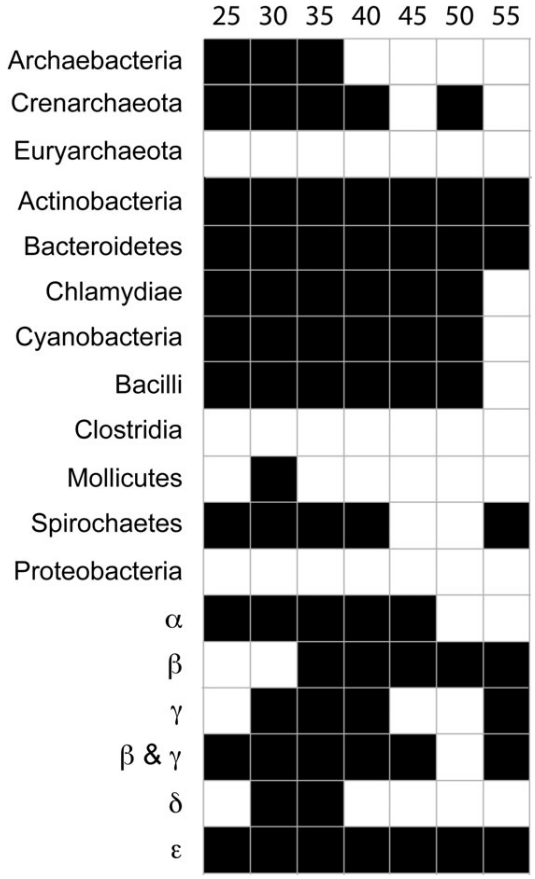


Fig. 6.—Detectable monophyly of groups under different similarity cutoffs, monophyly here meaning the presence of a split uniting the group irrespective of the presence of conflicting splits. Black square indicates that the respective group is monophyletic in that sense under the given cutoff.

Comparison of Evolutionary Rates among Lineages

Ancient paralogs are protein pairs that were duplicated prior to the divergence of eubacteria and archaebacteria (fig. 2*a*). Here, we use eight such paralogs in order to compare evolutionary rates among prokaryotic lineages (Kollman and Doolittle 2000): 1) adenosine triphosphate (ATP) synthase α (atpA) and β (atpB) subunits, 2) carbamoyl-phosphate synthase small (carA) and large (carB) subunits, 3) SRP proteins (ftsY, ffh), 4) isoleucyl-tRNA synthetase (ileS) and valyl-tRNA synthetase (valS), 5) aspartate carbamoyltransferase (pyrB) and ornithine carbamoyltransferase (argF), 6) threonyl-tRNA synthetase (thrS) and seryl-tRNA synthetase (serS), 7) translation elongation factors EF-G (fusA) and EF-Tu (tufA), and 8) tyrosyl-tRNA synthetase (tyrS) and tryptophanyl-tRNA synthetase (trpS). For all possible species pairs that represent two different higher taxa (called here lineages for convenience) shown in supplementary table S4 (Supplementary Material online), we investigated the corresponding ancient paralog quartet. Of course, LGT of ancient paralogs can generate topologies other than that expected by vertical inheritance alone (Zhaxybayeva et al. 2005). We therefore tested each quartet for a vertical topology (fig. 2*b*) using the SH test. Quartets of vertical topology (t_{vert}) were then tested for a global clock model (r_{global}) using the maximum-likelihood ratio test (Yang 2007). In the cases where r_{global} was rejected, the quartet was tested for lineage-specific rates (fig. 2*c*). Quartets of vertical topology that accepted the lineage-specific rates model ($r_{lineage}$) permit identification of lineage-specific rate increases, that is, which of the two genomes is undergoing more rapid sequence change.

Thus, orthologs of the eight ancient paralogs were identified in all genomes and were used to assemble 115,750 sequence quartet alignments. Alignment quality was tested using the HoT procedure (Landan and Graur 2007). Employing a conservative cutoff for alignment reliability of identical sum-of-pairs score >80% resulted in 56,297 alignments for which we reconstructed maximum-likelihood trees; the remaining 59,453 alignments were excluded because about half (49 ± 16%) of the site patterns (columns) in the alignment were irreproducible in the simplest alignment comparison (N-terminal vs. C-terminal seeding). The proportion of t_{vert} trees within consistent alignments is very high, ranging between 90% of ATP synthase quartets and 100% of the carbamoyl-phosphate synthetase and translation elongation factor EF-Tu and EF-G quartets (supplementary table S5, Supplementary Material online). In total, of the 56,297 reproducible alignments, 55,765 (99%) gave a t_{vert} quartet result. This high proportion of vertical topologies—for the paralogous two taxon case—suggests that LGT of these genes between the higher level taxonomic groups sampled here is quite rare, whereby this result does not address the frequency of transfer of these genes among closely related lineages. Using a maximum-likelihood ratio test, we were able to accept a global clock model for most (75%) of the t_{vert} quartets. Furthermore, 58% (5,611) of the quartets comparing archaebacterial and eubacterial lineages passed the global clock model test. Hence, in most cases, there is no significant difference in evolutionary rates between the different lineages for the proteins we tested (Novichkov et al. 2004).

We performed this test specifically to address the empirical validity of repeated assertions that the archaebacteria are an evolutionarily young group of organisms—only 850 million (Cavalier-Smith 2006a, 2009, 2010a, 2010b) or 1 billion (de Duve 2007) years of age—whose distinctness at the molecular level is attributable to some unspecified mutational mechanism of increased sequence change, quantum evolution (Cavalier-Smith 2010b), within the genome of archaebacteria in general or the archaebacterial common ancestor. Our results clearly indicate that there is no such lineage-specific effect for the archaebacteria (supplementary table S4, Supplementary Material online), although lineage-specific effects can be detected for other groups.

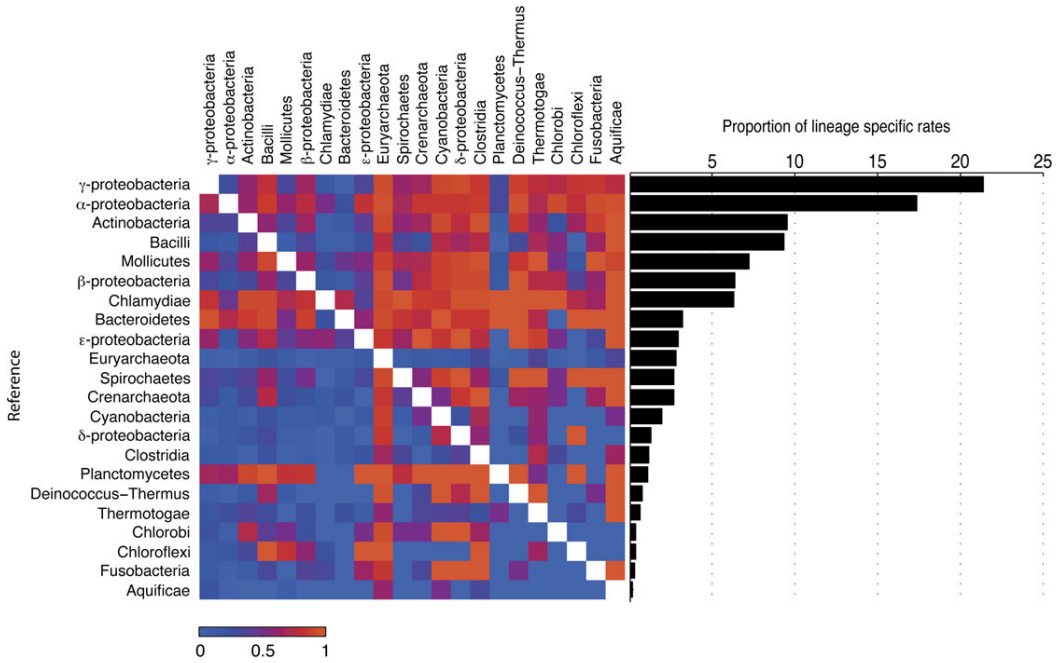


Fig. 7.—Rate comparisons of $r_{lineage}$ quartets. (*a*) A color-coded matrix showing the proportion of $r_{lineage}$ quartets in which the reference taxon (left) evolves in higher evolutionary rate than the compared taxon (top) in a 100% (red) to 0% (blue) scale. (*b*) Proportion of $r_{lineage}$ with elevated rates in the reference species from the total $r_{lineage}$ quartets in which the taxon is represented.

About one fifth (19%) of the total t_{vert} quartets uncover significant lineage-specific rate increases ($r_{lineage}$; supplementary table S4, Supplementary Material online); in these cases, both paralogs from the same lineage have the same degree of increased rate. Using these 10,728 $r_{lineage}$ quartets, we can compare the rates among lineages and rank lineages into slow- versus fast-evolving categories. The fastest lineages in this ranking are the γ-Proteobacteria, the α-Proteobacteria, the Actinobacteria, and the Bacilli (fig. 7). The splits of these four lineages within the splits networks are furthermore distinct across most protein similarity thresholds (fig. 3), suggesting a slight bias in the eubacterial clustering due to infraeubacterial evolutionary rate variation. But the two archaebacterial classes, euryarchaeota and crenarchaeota, are found to have at best an average rate in the lineage comparisons. They are slower than most eubacterial classes in the pairwise comparison (fig. 7), with only 4% (euryarchaeotes) and 10% (crenarchaeotes) of the t_{vert} quartets suggesting a higher rate in the respective archaebacterial class. Hence, the weight of the rooted split between archaebacteria and eubacteria cannot be attributed to faster archaebacterial evolutionary rates. Furthermore, the argument that archaebacteria are only 850–1,000 MY old (Cavalier-Smith 2006a; de Duve 2007) is rejected because its corollary that their molecular distinctness can be explained away by assuming an increased archaebacterial evolutionary rate is shown here to be untrue. Our findings are, however, fully consistent with the view that the archaebacteria are a very ancient lineage of organisms, at least as ancient as the eubacteria (Stetter 2006; Thauer 2007), a view that is furthermore consistent with isotope data for the antiquity of archaebacterial metabolism.

Life at the Root The debate about the position of the root in the tree of life has focused mainly on its position and to some extent on the biology of the first organisms. The issues of microbial lifestyle (autotrophy vs. heterotrophy: Lane et al. 2010) and cellularity, that is, the transition from replicating molecules in inorganic compartments to genetically specified replicating cells (Martin and Russell 2003; Koonin and Martin 2005; Branciamore et al. 2009) have received attention of late. However, by far the most heavily debated aspect of life at the root concerns temperature.

The view of thermophilic origins attracted much attention following the suggestions by Karl Stetter (Stetter et al. 1990) and Pace (1991) that prokaryotes inhabiting many of the extreme kinds of environments we see today are, to some extent, inhabiting environments that existed in a fully "modern" form on early earth: anoxic volcanic settings and hydrothermal vents, both which are often quite hot (>80 °C). In trees rooted between the prokaryotic domains, the hyperthermophiles branched first, suggesting that maybe the first organisms were hyperthermophilic archaea and bacteria (Stetter et al. 1990; Pace 1991). That view spawned the counterhypothesis of thermoreduction (Forterre 1995, 1996), which posits that the hyperthermophilic

origins scenario is wrong by virtue of a misplaced root. In that view, the eukaryotes are seen as the ancestral form of life on earth, prokaryotes having evolved from them via reductive evolution. Although thermoreduction in the original sense can now be excluded because all eukaryotes either have or had mitochondria (Cox et al. 2008; van der Giezen 2009), meaning that eukaryotes as we know them cannot be ancestral to prokaryotes, the issue of temperature at life's root remains current.

Recently, gene trees have been used to infer the temperature of early earth environments based on statistical arguments (Gaucher et al. 2003, 2008). Boussau et al. (2008), for example, suggested that the first organisms (the common ancestor of archaeabacteria and eubacteria in their view) arose and lived at about room temperature (~20 °C) based on the estimated GC content of inferred ancestral sequences in maximum-likelihood trees. Is such a low temperature for life at the root realistic? Amend and McCollom (2009) recently calculated that in geochemically promising environments for the origin of life, the Gibbs energy of reaction (ΔG_r) toward the synthesis of total prokaryotic cell mass was unfavorable (+500 Joules per gram of cells) at 25 °C but exergonic at 50, 75, and 100 °C, with values of –1,016, –873, and –628 Joules per gram of cells, respectively, dropping sharply again at 125 °C (Amend and McCollom 2009). Clearly, the synthesis of the first cells must have entailed a fundamentally exergonic reaction, as life cannot have arisen against the laws of thermodynamics. If thermodynamics are favorable in the range of 50–100 °C but not at 25 °C, then this can be taken as a constraint for phylogenetic models rather than a variable for estimation, when it comes to considering temperature at the root.

Part of the rational against the view of thermophilic origins was once founded in the circumstance that nucleoside triphosphates are very unstable at temperatures around 100 °C (Forterre 1996), for which reason such temperatures were deemed to be incompatible with the notion of an RNA world. However, Constanzo et al. (2009) recently reported that RNA chains dozens to over 100 nucleotides in length arise spontaneously, in hot (>80 °C) water, and without catalysts yet not from nucleoside triphosphates rather from the ribonucleoside 3′,5′ cyclic monophosphates at concentrations around 1 mM. Temperatures around 85 °C yielded rapid polymerization, below 60 °C the reaction rates dropped sharply (Constanzo et al. 2009). Thus, from the thermodynamic and chemical perspective, life at the root might be more likely in the range of 50–100 °C than at values approaching room temperature. That view is consistent with the recent discovery of a novel bifunctional fructose-1,6-bisphosphate aldolase/phosphatase from thermophilic eubacteria and archaebacteria that provides comparative biochemical evidence in favor of chemolithoautotrophic origins (Say and Fuchs 2010).

Conclusions

Recent studies on the position of the root of prokaryotic life have suggested that it lies within anoxygenic photosynthetic eubacteria (Cavalier-Smith 2006b) or within the eubacteria between the actinobacteria and the firmicutes (Lake et al. 2009). In such eubacterial root scenarios, the archaebacteria are seen as derived from specific groups of the eubacteria, in which case an elevated rate must be invoked for the archaebacteria in order to account for their molecular divergence. We have shown that no indication of such an archeabacterial rate elevation exists in available genome sequence data. Our analyses indicate that the deepest divide in the living world is that between archaebacteria and eubacteria, as earlier studies indicated (Gogarten et al. 1989; Iwabe et al. 1989) and as is compatible with much recent genome data (Koonin 2009). Like supertree approaches (Pisani et al. 2007), our method takes the signal of all genes—including those that have undergone LGT—into account rather than demanding that gene families harboring LGT events first be identified and purged from the data. In contrast to supertree and supermatrix methods, however, our procedure is independent of individual phylogenetic trees and utilizes an approach entailing phylogenetic networks to the study of evolutionary genome comparisons.

Supplementary Material

Supplementary figure S1 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft and the European Research Council for financial support. D.B. was supported by the Alexander von Humboldt Foundation and by a Marsden grant through the Royal Society of New Zealand.

Literature Cited

Amend JP, McCollom TM. 2009. Energetics of biomolecule synthesis on early Earth. In: Zaikowski L, Friedrich JM, Seidel SR, editors. Chemical evolution II: from the origins of life to modern society. Washington, DC: American Chemical Society. pp. 63–94.

Bapteste E, Philippe H. 2002. The potential value of indels as phylogenetic markers: position of trichomonads as a case study. Mol Biol Evol. 19:972–977.

Bapteste E, et al. 2009. Prokaryotic evolution and the tree of life are two different things. Biol Direct. 4:34.

Battistuzzi FU, Hedges SB. 2009. A major clade of prokaryotes with ancient adaptations to life on land. Mol Biol Evol. 26:335–343.

Bell SD, Jackson SP. 2001. Mechanism and regulation of transcription in archaea. Curr Opin Microbiol. 4:208–213.

Biegel E, Schmidt S, Muller V. 2009. Genetic, immunological and biochemical evidence for a Rnf complex in the acetogen *Acetobacterium woodii*. Environ Microbiol. 11:1438–1443.

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

Branciamore S, Gallori E, Szathmary E, Czaran T. 2009. The origin of life: chemical evolution of a metabolic system in a mineral honeycomb? J Mol Evol. 69:458–469.

Brasier MD, Mcloughlin N, Green O, Wacey D. 2006. A fresh look at the fossil evidence for early archaean cellular life. Philos Trans R Soc Lond B Biol Sci. 361:887–902.

Brown JR, Doolittle WF. 1995. Root of the universal tree of life based on ancient aminoacyl-transfer-RNA synthetase gene duplications. Proc Natl Acad Sci U S A. 92:2441–2445.

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21:255–265.

Butterfield NJ. 2000. *Bangiomorpha pubescens* n. Gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. Paleobiology. 26:386–404.

Canfield DE. 2006. Biogeochemistry—gas with an ancient history. Nature. 440:426–427.

Cavalier-Smith T. 2006a. Cell evolution and earth history: stasis and revolution. Philos Trans R Soc Lond B Biol Sci. 361:969–1006.

Cavalier-Smith T. 2006b. Rooting the tree of life by transition analyses. Biol Direct. 1:19.

Cavalier-Smith T. 2009. Predation and eukaryote cell origins: a co-evolutionary perspective. Int J Biochem Cell Biol. 41:307–322.

Cavalier-Smith T. 2010a. Deep phylogeny, ancestral groups and the four ages of life. Philos Trans R Soc Lond B Biol Sci. 365:111–132.

Cavalier-Smith T. 2010b. Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution. Biol Direct. 5:7.

Chistoserdova L, Vorholt JA, Thauer RK, Lidstrom ME. 1998. C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. Science. 281:99–102.

Chong JPJ, Hayashi MK, Simon MN, Xu RM, Stillman B. 2000. A double-hexamer archaeal minichromosome maintenance protein is an ATP-dependent DNA helicase. Proc Natl Acad Sci U S A. 97:1530–1535.

Claus H, et al. 2005. Molecular organization of selected prokaryotic S-layer proteins. Can J Microbiol. 51:731–743.

Constanzo G, Pino S, Ciciriello F, Di Mauro E. 2009. Generation of long RNA chains in water. J Biol Chem. 284:33206–33216.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A. 105:20356–20361.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105:10039–10044.

Dagan T, Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci U S A. 104:870–875.

de Duve C. 2007. The origin of eukaryotes: a reappraisal. Nat Rev Genet. 8:395–403.

Deppenmeier U. 2002. The unique biochemistry of methanogenesis. Prog Nucleic Acid Res Mol Biol. 71:223–283.

Deppenmeier U, et al. 2002. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol. 4:453–461.

Di Giulio M. 2007. The evidence that the tree of life is not rooted within the Archaea is unreliable: a reply to Skophammer et al. 2007. Gene. 394:105–106.

Dimarco AA, Bobik TA, Wolfe RS. 1990. Unusual coenzymes of methanogenesis. Annu Rev Biochem. 59:355–394.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science. 284:2124–2128.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci U S A. 104:2043–2049.

Dress AW, Huson DH. 2004. Constructing splits graphs. IEEE/ACM Trans Comput Biol Bioinform. 1:109–115.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. Nature. 440:623–630.

Engelhardt H. 2007. Are S-layers exoskeletons? The basic function of protein surface layers revisited. J Struct Biol. 160:115–124.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Farris JS. 1972. Estimating phylogenetic trees from distance matrices. Am Nat. 106:645–668.

Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.

Fischer WW. 2008. Biogeochemistry—life before the rise of oxygen. Nature. 455:1051–1052.

Forterre P. 1995. Thermoreduction: a hypothesis for the origin of prokaryotes. C R Acad Sci III. 318:415–422.

Forterre P. 1996. A hot topic: the origin of hyperthermophiles. Cell. 85:789–792.

Frols S, White MF, Schleper C. 2009. Reactions to UV damage in the model archaeon sulfolobus solfataricus. Biochem Soc Trans. 37:36–41.

Fujihashi M, et al. 2007. Crystal structure of archaeal photolyase from *Sulfolobus tokodaii* with two FAD molecules: implication of a novel light-harvesting cofactor. J Mol Biol. 365:903–910.

Fukami-Kobayashi K, Minezaki Y, Tateno Y, Nishikawa K. 2007. A tree of life based on protein domain organizations. Mol Biol Evol. 24:1181–1189.

Gaucher EA, Govindarajan S, Ganesh OK. 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature. 451:704–707.

Gaucher EA, Thomson JM, Burgan MF, Benner SA. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature. 425:285–288.

Gogarten JP, et al. 1989. Evolution of the vacuolar H+-ATPase—implications for the origin of eukaryotes. Proc Natl Acad Sci U S A. 86:6661–6665.

Graham DE, Overbeek R, Olsen GJ, Woese CR. 2000. An archaeal genomic signature. Proc Natl Acad Sci U S A. 97:3304–3308.

Grassineau NV, Abell P, Appel PWU, Lowry D, Nisbet EG. 2006. Early life signatures in sulfur and carbon isotopes from Isua, Barberton, Wabigoon (Steep Rock), and Belingwe greenstone belts (3.8 to 2.7 Ga). Geol Soc Am Mem. 198:33–52.

Gupta RS. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev. 62:1435–1491.

Gupta RS, Lorenzini E. 2007. Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the Bacteroidetes and Chlorobi species. BMC Evol Biol. 7:71.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci U S A. 86:9355–9359.

Javaux EJ, Knoll AH, Walter MR. 2001. Morphological and ecological complexity in early eukaryotic ecosystems. Nature. 412:66–69.

Genome Biology and Evolution

SMBE

Knoll AH, Javaux EJ, Hewitt D, Cohen P. 2006. Eukaryotic organisms in Proterozoic oceans. Philos Trans R Soc Lond B Biol Sci. 361:1023–1038.
Kollman JM, Doolittle RF. 2000. Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. J Mol Evol. 51:173–181.
Koonin EV. 2009. Darwinian evolution in the light of genomics. Nucleic Acids Res. 37:1011–1034.
Koonin EV, Martin W. 2005. On the origin of genomes and cells within inorganic compartments. Trends Genet. 21:647–654.
Lake JA, Herbold CW, Rivera MC, Servin JA, Skophammer RG. 2007. Rooting the tree of life using nonubiquitous genes. Mol Biol Evol. 24:130–136.
Lake JA, Servin JA, Herbold CW, Skophammer RG. 2008. Evidence for a new root of the tree of life. Syst Biol. 57:835–843.
Lake JA, Skophammer RG, Herbold CW, Servin JA. 2009. Genome beginnings: rooting the tree of life. Philos Trans R Soc Lond B Biol Sci. 364:2177–2185.
Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.
Lane N, Allen JF, Martin W. 2010. How did LUCA make a living? Chemiosmosis and the origin of life. Bioessays. 32:271–280.
Lang AS, Beatty JT. 2007. Importance of widespread gene transfer agent genes in alpha-proteobacteria. Trends Microbiol. 15:54–62.
Ljungdahl LG. 2009. A life with acetogens, thermophiles, and cellulolytic anaerobes. Annu Rev Microbiol. 63:1–25.
Martin HH, König H. 1996. Beta-lactamases are absent from archaea (archaebacteria). Microb Drug Resist. 2:269–272.
Martin W, Baross J, Kelley D, Russell MJ. 2008. Hydrothermal vents and the origin of life. Nat Rev Microbiol. 6:805–814.
Martin W, Russell M. 2003. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Philos Trans R Soc Lond B Biol Sci. 358:59–85.
McInerney JO, Cotton JA, Pisani D. 2008. The prokaryotic tree of life: past, present... and future? Trends Ecol Evol. 23:276–281.
McInerney JO, Pisani D. 2007. Genetics—paradigm for life. Science. 318:1390–1391.
Müller V. 2003. Energy conservation in acetogenic bacteria. Appl Environ Microbiol. 69:6345–6353.
Ng WV, et al. 2000. Genome sequence of Halobacterium species NRC-1. Proc Natl Acad Sci U S A. 97:12176–12181.
Nisbet E. 2000. Palaeobiology: the realms of Archaean life. Nature. 405:625–626.
Novichkov PS, et al. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J Bacteriol. 186:6575–6585.
Pace NR. 1991. Origin of life: facing up to the physical setting. Cell. 65:531–533.
Pierce E, et al. 2008. The complete genome sequence of *Moorella thermoacetica* (f. *Clostridium thermoaceticum*). Environ Microbiol. 10:2550–2573.
Pisani D, Cotton JA, Mcinerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. Mol Biol Evol. 24:1752–1760.
Puigbò P, Wolf YI, Koonin EV. 2009. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J Biol. 8:59.
Raoult D, et al. 2003. *Tropheryma whipplei* Twist: a human pathogenic Actinobacteria with a reduced genome. Genome Res. 13:1800–1809.
Rasmussen B. 2000. Filamentous microfossils in a 3,235-million-year-old volcanogenic massive sulphide deposit. Nature. 405:676–679.
Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR. 2008. Reassessing the first appearance of eukaryotes and cyanobacteria. Nature. 455:1101–1104.
Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature. 431:152–155.
Saitou N, Nei M. 1987. The Neighbor-Joining method: a new method for reconstruction of phylogenetic trees. Mol Biol Evol. 4:406–425.
Say RF, Fuchs G. 2010. Fructose-1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. Nature. 464:1077–1081.
Shen Y, Buick R, Canfield DE. 2001. Isotopic evidence for microbial sulphate reduction in the early Archaean era. Nature. 410:77–81.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.
Skophammer RG, Herbold CW, Rivera MC, Servin JA, Lake JA. 2006. Evidence that the root of the tree of life is not within the Archaea. Mol Biol Evol. 23:1648–1651.
Sleep NH, Meibom A, Fridriksson T, Coleman RG, Bird DK. 2004. H_2-rich fluids from serpentinization: geochemical and biotic implications. Proc Natl Acad Sci U S A. 101:12818–12823.
Snel B, Bork P, Huynen MA. 1999. Genome phylogeny based on gene content. Nat Genet. 21:108–110.
Stetter KO. 2006. Hyperthermophiles in the history of life. Philos Trans R Soc Lond B Biol Sci. 361:1837–1842.
Stetter KO, Fiala G, Huber G, Huber R, Segerer A. 1990. Hyperthermophilic microorganisms. FEMS Microbiol Rev. 75:117–124.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278:631–637.
Thauer RK. 2007. A fifth pathway of carbon fixation. Science. 318:1732–1733.
Thauer RK, Kaster AK, Seedorf H, Buckel W, Hedderich R. 2008. Methanogenic archaea: ecologically relevant differences in energy conservation. Nature Rev Microbiol. 6:579–591.
Thompson JD, Higgins DG, Gibson TJ. 1994. ClustalW—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Ueno Y, Yamada K, Yoshida N, Maruyama S, Isozaki Y. 2006. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. Nature. 440:516–519.
van der Giezen M. 2009. Hydrogenosomes and mitosomes: conservation and evolution of functions. J Eukaryot Microbiol. 56:221–231.
Ventura GT, et al. 2007. Molecular evidence of Late Archean archaea and the presence of a subsurface hydrothermal biosphere. Proc Natl Acad Sci U S A. 104:14260–14265.
Wong JT, Chen J, Mat WK, Ng SK, Xue H. 2007. Polyphasic evidence delineating the root of life and roots of biological domains. Gene. 403:39–52.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.
Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol. 17:1081–1090.
Zhaxybayeva O, Lapierre P, Gogarten JP. 2005. Ancient gene duplications and the root(s) of the tree of life. Protoplasma. 227:53–64.
Zhaxybayeva O, et al. 2009. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. Proc Natl Acad Sci U S A. 106:5865–5870.

Associate editor: Eugene Koonin

5.2 A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies

Mayo Roettger, William Martin, and Tal Dagan

Institut für Botanik III, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

Der vorgestellte Artikel wurde in der Fachzeitschrift *Molecular Biology and Evolution* veröffentlicht (Roettger, Martin und Dagan, 2009).

Beitrag von Mayo Röttger:

Versuchsplanung:	50 %
Datenanalyse:	90 %
Verfassen des Manuskripts:	30 %

RESEARCH ARTICLES

A Machine-Learning Approach Reveals That Alignment Properties Alone Can Accurately Predict Inference of Lateral Gene Transfer from Discordant Phylogenies

Mayo Roettger, William Martin, and Tal Dagan

Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Germany

Among the methods currently used in phylogenomic practice to detect the presence of lateral gene transfer (LGT), one of the most frequently employed is the comparison of gene tree topologies for different genes. In cases where the phylogenies for different genes are incompatible, or discordant, for well-supported branches there are three simple interpretations for the result: 1) gene duplications (paralogy) followed by many independent gene losses have occurred, 2) LGT has occurred, or 3) the phylogeny is well supported but for reasons unknown is nonetheless incorrect. Here, we focus on the third possibility by examining the properties of 22,437 published multiple sequence alignments, the Bayesian maximum likelihood trees for which either do or do not suggest the occurrence of LGT by the criterion of discordant branches. The alignments that produce discordant phylogenies differ significantly in several salient alignment properties from those that do not. Using a support vector machine, we were able to predict the inference of discordant tree topologies with up to 80% accuracy from alignment properties alone.

Introduction

The phylogenetic approach for lateral gene transfer (LGT) inference from the frequency of incongruent branching patterns in gene trees has so far delivered widely conflicting results, ranging from estimates that as few as 2% (Ge et al. 2005) to possibly 14% of all genes in prokaryote genomes are affected by LGT (Beiko, Harlow, and Ragan 2005). Such divergent estimates using phylogenetic tree comparisons can, in principle, be attributed to many factors including the obvious, such as lineage sampling, the inherent uncertainties of various approaches to phylogenetic reconstruction (Penny et al. 1992; Hillis 1995; Lopez et al. 2002) and the threshold levels of support set to score the presence of genuinely conflicting topologies. But phylogenetic trees of molecular sequences are always inferred from multiple sequence alignments. Nei et al. (1995) and Nei (1996) pointed out early on that alignment of highly diverged sequences may result in erroneous phylogenetic reconstruction. Interest in this aspect of phylogeny has renewed with several reports investigating the alignment step itself as it specifically relates to phylogenetic inference (Landan and Graur 2007; Deusch et al. 2008; Löytynoja and Goldman 2008; Wong et al. 2008)

Here, we wished to examine the extent to which LGT inference by the phylogenetic method might be sensitive to the properties of alignments themselves. For this purpose, we investigated the comprehensive data set compiled and carefully analyzed by Beiko, Harlow, and Ragan (2005), who kindly made their data available. Their data set is highly suitable for the present study 1) because it consists of 22,437 carefully assembled gene families of prokaryotic orthologs, in which paralogs have been sorted out by using a conservative similarity cutoff (Beiko, Harlow, and Ragan 2005, Supplementary Material online), 2) because they used a widely employed filter, Gblocks (Castresana 2000), to exclude poorly aligned regions from their analysis prior to phylogenetic reconstruction, and 3) because they used a very stringent (conservative) threshold for the scoring of discordant phylogenies. In brief, Beiko, Harlow, and Ragan (2005) constructed a consensus supertree for the proteins encoded in 144 prokaryotic genomes and constructed from the same data 22,437 individual phylogenetic trees containing from 4 to 144 sequences each using a Bayesian approach. They inferred LGT only from highly significant (posterior probability ≥ 0.95) discordant tree topologies in comparison to the consensus supertree topology (Beiko, Harlow, and Ragan 2005). For 5,822 of those trees, one or more LGT was inferred on the basis of discordance to the consensus topology, we designate those amino acid sequence alignments as "LGT positive" or LGT for short. The remaining 16,615 of the alignments investigated by Beiko, Harlow, and Ragan (2005) did not produce branches (bipartitions) that were discordant (conflicting) with the consensus supertree topology and are considered here as "vertical gene inheritance" or VGI alignments. We examined the properties of the LGT alignments in comparison to the properties of the VGI alignments.

Methods

For the analysis, we used a data set of 22,437 protein families from 144 prokaryotes for which LGT) was inferred using the phylogenetic method (Beiko, Harlow, and Ragan 2005). The data for each protein family include a multiple sequence alignment yielding the highest score according to the word-oriented objective function (Beiko, Chan, and Ragan 2005) from a set of alignments reconstructed by several different algorithms: ClustalW (Thompson et al. 1994), T-coffee (Notredame et al. 2000), MAFFT (Katoh et al. 2002), POA (Grasso and Lee 2004), and PRRP (Gotoh 1996), a partial alignment of relatively conserved regions constructed with Gblocks (Castresana 2000), and a phylogenetic tree inferred with MrBayes (Huelsenbeck and Ronquist 2001). Bipartitions

Key words: lateral gene transfer, molecular phylogeny, discordant tree topologies, support vector machine, principal component analysis.

E-mail: mayo.roettger@uni-duesseldorf.de.

Mol. Biol. Evol. 26(9):1931–1939. 2009
doi:10.1093/molbev/msp105
Advance Access publication May 14, 2009

in the phylogenetic tree were considered as concordant if they overlap with the reference supertree, or discordant otherwise, which were interpreted as LGT events (Beiko, Harlow, and Ragan 2005).

Multiple Alignment Properties

For each protein family (alignment), we calculated alignment properties as follows: Number of operational taxonomic units (OTUs) is the number of orthologs in the family. Proportion of gaps is the proportion of gap characters in the Gblocks output alignment. Entropy was calculated for each Gblocks output alignment as the average entropy of its sites. For the calculation, we used the Shannon information content normalized by the number of OTUs in the alignment (Valdar 2002):

$$\text{Entropy} = \frac{\sum_{\text{col}=1}^{N_{\text{sites}}}\left(-\lambda_t \sum_{i=1}^{K} p_{i,\text{col}}\log_2 p_{i,\text{col}}\right)}{N_{\text{sites}}},$$

where N_{sites} is the number of alignment sites, K is the alphabet size, which is 21 in this case (20 amino acids plus one gap symbol), and $p_{i,\text{col}}$ the probability of observing the ith character in alignment column col. The λ_t factor is used to scale the entropy into a [0,1] range by the number of OTUs (N_{seq}) and the alphabet size K.

$$\lambda_t = \left[\log_2(\min(N_{\text{seq}}, K))\right]^{-1}.$$

Invariant sites are positions in the Gblocks alignment where all sequences contain the same amino acid, and informative sites are defined as alignment columns containing at least two different amino acids, each one observed in at least two sequences at the position. Average pairwise identity is calculated as the proportion of amino acid identities between all sequence pairs and averaged for the protein family as follows:

$$\text{API} = \frac{2}{N_{\text{seq}}(N_{\text{seq}}-1)N_{\text{sites}}} \sum_{\text{col}=1}^{N_{\text{sites}}} \sum_{i=1}^{N_{\text{seq}}-1} \sum_{j=i+1}^{N_{\text{seq}}} \left\{ \begin{array}{lll} 1 : a_{i,\text{col}} & = & a_{j,\text{col}} \\ 0 : a_{i,\text{col}} & \neq & a_{j,\text{col}} \end{array} \right\},$$

where $a_{i,\text{col}}$ is the observed amino acid in Gblocks-alignment sequence i at position col.

In addition, we tested for alignment reliability using the Heads-or-Tails (HoT) method (Landan and Graur 2007). Tails alignments were obtained by aligning the reversed sequences of each protein family using exactly the same alignment procedure that was used for the original (Heads) alignments. Column score (CS) is calculated as the proportion of identical columns between Heads and Tails alignments, and sum of pairs score (SPS) was calculated as the proportion of identical residue position pairs between Heads and Tails alignments.

Each alignment comprises protein sequences from different species, each sequence named by Beiko et al. with a unique pipeline id. Additional files contained information about the original gi number in the RefSeq database (Pruitt et al. 2005) together with the current gi number in the database and the genome id of the sequence used by the National Center for Biotechnology Information (NCBI). Two hundred and twenty-three proteins in 183 alignments had no information about the sequence except the original gi number in the database. These database entries were replaced or removed from the database. In the calculation of the number of different phyla and the classification of the sequence into the kingdom groups, we discarded these sequences from the alignment. We obtained taxonomical classification information for each sequence from NCBI and counted the number of sequences being classified as different phyla for each cluster. We used the term archaea for clusters that contain only sequences classified as of archaebacterial origin, the term eubacteria for clusters that contain only sequences classified as of bacterial origin, and the term universal for clusters that contain sequences of both kingdoms.

For the comparison of the property distributions between LGT and VGI alignments, the Wilcoxon nonparametric test was used.

Orthologs Pairwise Distances

Protein pairwise distances between orthologs from LGT and VGI families were calculated for several genome pairs that were selected for their high frequency in the data: 1) *Vibrio vulnificus* versus *Yersinia pestis* (1,406 protein pairs where both species were present in the respective protein families), 2) *Brucella suis* versus *Mesorhizobium loti* (1,697 protein pairs), 3) *Agrobacterium tumefaciens* versus *M. loti* (2,205 protein pairs), and 4) *Bradyrhizobium japonicum* versus *M. loti* (1,794 protein pairs), 5) *Staphylococcus aureus* versus *Bacillus cereus* (924 protein pairs), 6) *Nostoc* sp. versus *Pyrococcus furiosus* (114 protein pairs), and 7) *Bacteroides thetaiotaomicron* versus *Sulfolobus solfataricus* (62 protein pairs). Pairwise protein distances were extracted from the distance matrix calculated from the multiple sequence alignments with PROTDIST (Felsenstein 1996) using Jonen-Taylor-Thornton substitution matrix (Jones et al. 1992). In addition, we calculated pairwise distances with the same method after realigning the orthologous sequences using MUSCLE (Edgar 2004).

Classification Procedure

Prediction of LGT (discordant tree bipartition) from alignment properties entailed a support vector machine (SVM) classifier (Christiani and Shawe-Taylor 2000). For the SVM training and classifying procedures, we used the svmtrain and svmclassify functions from the MATLAB 7.6 bioinformatics toolbox with the following parameters: Radial basis function (RBF) kernel, RBFSigmaValue = 1, Mlp_ParamsValue = [1,−1], MethodValue = SMO, BoxConstraintValue = 1, and AutoscaleValue = true. In order to obtain significance levels for the SVM performance, we applied 10-fold crossvalidation in each step using the small 1/10 subset for training and the 9/10 for testing. The LGT/VGI ratio in the training set was adjusted by randomly selecting different numbers of LGT and VGI samples from the preliminary training set to form an equal-sized training set for each validation step.

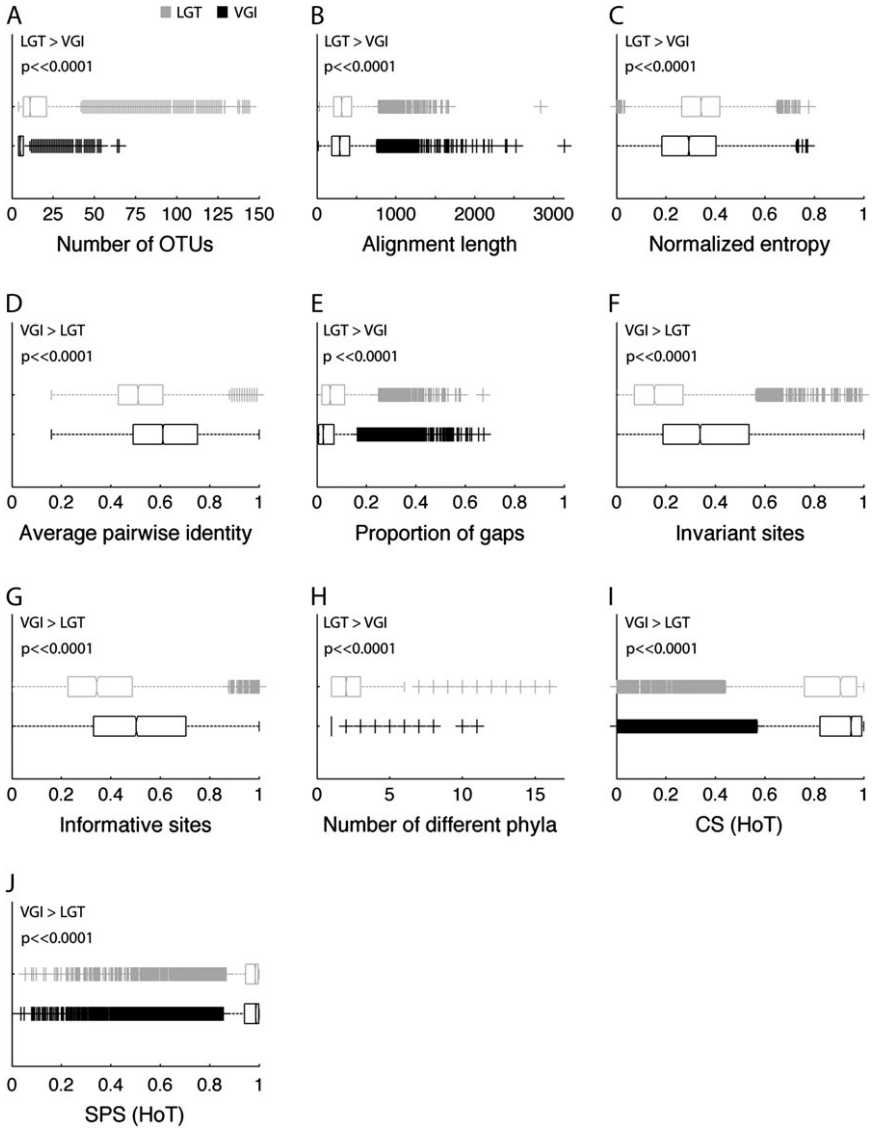


FIG. 1.—Distributions of alignment properties in the LGT and VGI groups. Differences in the distributions of the two groups were tested by the Wilcoxon nonparametric test (*P* values presented at the top of each graph).

SVM performance was evaluated by "accuracy," that is, the proportion of alignments correctly classified as LGT or VGI, "sensitivity," which is the true positive rate or the number of true positives (LGT alignments that are classified as such) divided by the sum of true positives plus false negatives (LGT alignments classified as VGI), and "specificity," as the true negative rate or the number of true negatives (VGI alignments that are classified as such) divided by the sum of true negatives plus false positives (VGI alignments classified as LGT).

To test the performance of the classifier under different LGT/VGI proportions in the training set and in the test set, we used LGT proportions ranging from 25% to 75% while including all 11 alignment properties.

To explore the contribution of the different features and their combinations to the classification performance, we tested all possible 2,047 combinations of the 11 alignment features analyzed in this study using a training set with equal proportions of LGT and VGI alignments.

Multivariate Analysis

We performed principal component analysis (PCA) using the princomp function of MATLAB 7.6. The data for each alignment property were normalized before the analysis, so that all properties had only values ranging from zero to one.

Results and Discussion

It is known that the probability of obtaining incorrect trees increases with the number of sequences (OTUs) analyzed (Nei 1996). The LGT alignments investigated here contained significantly larger numbers of OTUs ($P \ll 0.0001$) than the VGI alignments (fig. 1*A*). No VGI alignment in the present sample contains more than 65 sequences, whereas 5% of the LGT alignments contain $\geq$65 sequences. It is also known that for a given level of sequence divergence, the probability of obtaining incorrect trees is

Table 1
General Statistics of Protein Family Alignment Properties Grouped by LGT and VGI Categories

	Range (Min–Max)		Mean ± SD		Median	
MSA Parameter	VGI	LGT	VGI	LGT	VGI	LGT
Normalized Shannon entropy	0.000–0.772	0.002–0.777	0.294 ± 0.151	0.343 ± 0.117	0.292	0.341
Average pairwise identity	0.160–1.000	0.160–0.990	0.621 ± 0.176	0.523 ± 0.132	0.610	0.510
Proportion of gaps	0.000–0.675	0.000–0.671	0.052 ± 0.072	0.080 ± 0.084	0.025	0.053
Number of OTUs	4–65	4–144	6.6 ± 4.2	19.0 ± 21.7	5.0	11.0
Alignment length	14–3,135	31–2,837	325.0 ± 203.9	352.0 ± 212.3	287.0	311.0
Proportion of invariant sites	0.000–1.000	0.000–0.992	0.381 ± 0.247	0.191 ± 0.160	0.337	0.153
Proportion of informative sites	0.000–1.000	0.000–1.000	0.524 ± 0.250	0.366 ± 0.188	0.502	0.343
CS (HoT)	0.000–1.000	0.000–1.000	0.856 ± 0.212	0.823 ± 0.213	0.948	0.905
SPS (HoT)	0.037–1.000	0.053–1.000	0.939 ± 0.118	0.943 ± 0.111	0.986	0.984
Number of different phyla	1–11	1–16	1.323 ± 0.708	2.651 ± 2.475	1.0	2.0

NOTE.—The data set contains 22,437 protein family alignments, 5,822 of which are LGT and 16,615 are VGI.

higher when short sequences are analyzed than when longer sequences are analyzed (Nei 1996). However, the LGT protein families investigated here contain sequences that are significantly ($P < 0.0001$) longer than VGI protein families (supplementary fig. S1, Supplementary Material online), producing also longer alignments (mean = 352; table 1) than the VGI alignments (mean = 325; table 1) (fig. 1*B*), suggesting that if incorrect trees are involved in LGT inference in the present data, then short sequences are not the cause.

The probability of obtaining incorrect trees increases when sequence divergence becomes too great (Nei 1996). Several alignment properties can address the issue of sequence divergence. Normalized Shannon entropy provides an estimate for the average number of different amino acids that occur per site in an alignment (Valdar 2002). The mean normalized Shannon entropy of LGT alignments is about 17% higher than for the VGI alignments in the present data (fig. 1*C*), a highly significant difference ($P < 0.0001$). Average sequence identity across all pairwise comparisons is a very simple and robust measure of sequence variability in an alignment. The average pairwise identity of the VGI alignments (mean = 0.621; table 1) is significantly higher ($P < 0.0001$) than in the LGT alignments (mean = 0.523) (fig. 1*D*). In addition, LGT alignments contain on average 50% more gaps than VGI alignments (fig. 1*E*; table 1). Another proxy for sequence divergence in an alignment is the proportion of invariant sites, the mean of which is 2-fold higher ($P < 0.0001$) in the VGI alignments than in the LGT alignments (fig. 1*F*). Furthermore, the proportion of informative sites, defined here as alignment columns containing at least two different amino acids each observed in at least two sequences at the position, is significantly lower ($P < 0.0001$) in the LGT alignments (mean = 0.366; table 1) than in the VGI alignments (mean = 0.524; table 1) (fig. 1*G*).

Thus, several alignment parameters that are known to increase the probability of obtaining incorrect trees—higher numbers of OTUs, sequence divergence exceeding 50% differences on average, and low numbers of informative sites—are significantly different in the LGT and the VGI alignments, and in all cases, LGT alignments are skewed toward the value that increases the probability of obtaining an incorrect tree. This does not directly indicate that the LGT alignments have produced branches that are highly supported but nonetheless incorrect (Delsuc et al. 2003), yet the tendency is consistent.

The proportion of invariant sites, the proportion of informative sites, and average pairwise identity show an inverted trend to the Shannon entropy in the PCA of the total data set (fig. 2). These three measures correlate negatively with Shannon entropy ($r = -0.84$, $r = -0.82$, and $r = -0.97$, $P < 0.0001$, respectively, supplementary fig. S2A–C, Supplementary Material online). This means that the less variable alignments may lack phylogenetic information due to high proportions of invariable sites, where the proportion of informative sites in these alignments will still be high. Yet these correlation coefficients are weaker in the LGT alignments ($r = -0.70$ and $r = -0.77$ and $r = -0.95$, $P < 0.0001$, respectively, supplementary fig. S2D–F, Supplementary Material online), than in the VGI alignments ($r = -0.87$ and $r = -0.84$ and $r = -0.98$, $P < 0.0001$, respectively; supplementary fig. S2G–I, Supplementary Material online) so that even though the LGT alignments are more variable than the VGI alignments, they generally contain not only fewer invariant sites but also fewer informative sites.

High alignment variability in the LGT alignments could be also the result of large numbers of sequences per alignment, as is the case for the LGT group alignments (fig. 1*A*). However, we found no correlation between number of OTUs and normalized entropy ($r = 0.01$, $P = 0.27$), and only weak correlation between number of OTUs and average pairwise identities ($r = -0.11$, $P < 0.0001$), or the proportion of gaps ($r = 0.23$, $P < 0.0001$; supplementary fig. S3, Supplementary Material online). Also, the number of phyla represented in the alignment, another possible source for higher alignment variability, is higher in the LGT groups than in the VGI group (fig. 1*H*). But this measure as well shows no significant correlation with any of the variability measures (supplementary fig. S4, Supplementary Material online). Hence, the high variability of the LGT alignments is not explained by the large number of sequences or the large number of phyla represented in these families.

The more variable the sequences in an alignment are, the more difficult they are to align and the more likely it is that the alignment procedures themselves can produce collections of site patterns that induce topological effects at the tree-building stage (Landan and Graur 2007; Wong et al. 2008). Thus, the LGT alignments, which are more variable than those in the VGI group, might be more error prone at the alignment step than the VGI alignments. To estimate

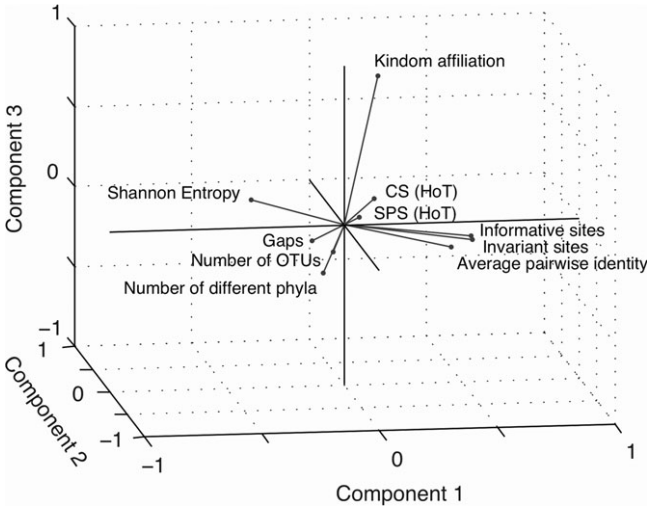


FIG. 2.—Principal component analysis of alignment properties. The axes represent the first three components, explaining 85% of the variability in the data (see supplementary tables S3 and S4, Supplementary Material online for details). Alignment properties are represented as vectors of their principal component coefficients. Alignment length is omitted due to its marginal contribution to the first three principal components. Two-dimensional views of every two respective components can be found in supplementary fig. S7, Supplementary Material online.

this effect, we compared alignment reliability of LGT and VGI alignments using the HoT method (Landan and Graur 2007). For these HoT comparisons, we realigned the original sequences (i.e., before filtering with Gblocks) as kindly provided by Beiko, Harlow, and Ragan (2005) in the C-to-N-direction to form the Tails alignments and compared them with the original (Heads) alignments. Both HoT parameters show an inverted trend to the number of OTUs, number of different phyla, and the proportion of gaps in the PCA analysis (fig. 2). Moreover, we found that the LGT alignments have a significantly ($P < 0.0001$) lower CS, which is the proportion of site columns reconstructed identically in the Heads and Tails alignment (fig. 1*I*), and a slightly but significantly ($P \ll 0.0001$) lower SPS, which is the proportion of identically reconstructed site pairs (fig. 1*J*), than the VGI alignments. Hence, LGT alignments contain significantly more alignment artifacts that are introduced by the sequence alignment process alone, independent of subsequent tree-building procedures. The bias in alignment quality within the LGT set is unlikely to be related with the erroneous guide tree used for the alignment because alignment errors are only marginally affected by the guide-tree quality (Landan and Graur 2008). Beiko, Harlow, and Ragan (2005) used a very conservative rule for inclusion in the LGT set that comprises only trees having at least one highly significant ("posterior probability" $\geq$ 0.95) discordant branch, whereas all other trees are considered as VGI. This results in abias toward highly supported (though not necessarily true) trees in the LGT set, where the proportion of highly significant branches per tree is 47 $\pm$ 28% versus 48 $\pm$ 40% (median 41% vs. 33%) in the VGI set. To test if this bias is related to the differences we found in the alignment properties, we deleted from the VGI set those alignments yielding trees with no highly significant branches, leaving 13,811 alignments yielding trees having at least one highly significant (posterior probability $\geq$ 0.95) concordant branch. This resulted in a set of trees, designated here VGI95, having a much higher proportion of highly significant branches per tree (57 $\pm$ 37%, median 50%). A comparison of alignment properties between the LGT and VGI95 sets resulted in identical conclusions to those detailed above for the comparison between the LGT and VGI sets (supplementary fig. S5, Supplementary Material online), so that the bias toward highly resolved trees in the LGT set has no relation to the bias in multiple alignment properties.

The comparison of alignment properties between VGI and LGT alignments summarized so far (fig. 1; table 1) shows that the LGT alignments are more variable than the VGI alignments. It is thus possible that the laterally transferred protein-coding sequences are inherently more variable than vertically inherited ones. To test this possibility, we compared pairwise protein distances between genomes to see if there were differences between LGT and VGI sequences with respect to overall sequence conservation. If so, then orthologous sequence pairs from VGI alignments should have smaller protein distances (i.e., should be more conserved) than orthologous pairs from LGT alignments. We tested that hypothesis for seven frequent genome pairs having proteins in both groups. The contrary was observed: Orthologous pairs from LGT alignments are more conserved (i.e., have smaller protein distances) than orthologous pairs from VGI alignments (fig. 3; supplementary fig. S6, Supplementary Material online). Hence, the higher variability observed in LGT alignments cannot be explained by a systematic bias in protein conservation among inherited versus laterally transferred proteins. This conclusion seems to contradict the bias toward variable multiple sequence alignments in the LGT set. A possible reconciliation between these two findings may be found in a study by Elhaik et al. (2006) showing that conserved proteins have higher probability of being detected by a similarity search, which leads to the composition of larger protein families, hence alignments with more OTUs that are probably more difficult to align. However, we found no correlation between the number of OTUs and the different alignment variability measures in our data set (supplementary fig. S3, Supplementary Material online).

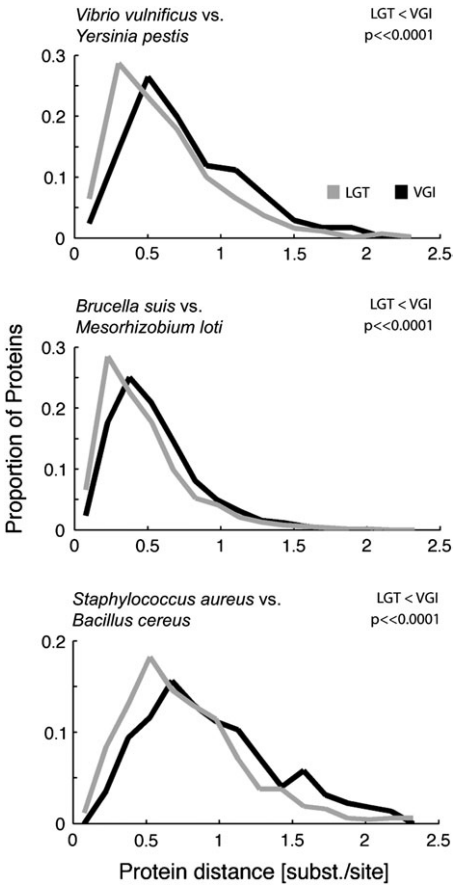


FIG. 3.—Comparison of protein pairwise distances between genome pairs found in LGT and VGI families. Distance distributions were compared using the Wilcoxon nonparametric test (*P* values presented at the top right of the graph).

Clustering of conserved orthologous proteins results not only in bigger families, but also in families having proteins from many taxomonic groups (supplementary fig. S4A, Supplementary Material online). However, the number of phyla alone is unlikely to reflect the relatedness among the sequences in the protein family because a protein family including sequences from eubacteria and archaebacteria is expected to contain more variability than a protein family including sequences from eubacteria only. Therefore, we divided the multiple alignments into those that comprise 1) eubacterial proteins only, 2) archaeal proteins only, and 3) "universal" alignments including proteins from both groups. A comparison of alignment properties among these three categories shows that universal families are much bigger than either eubacterial or archaeal families (fig. 4*A*). Moreover, all variability measures show that the universal alignments are more variable than alignments in the other two categories: Their entropy is higher (fig. 4*C*), their mean pairwise distance is higher (fig. 4*D*), and they contain more gaps (fig. 4*E*) and less invariant sites (fig. 4*F*). Universal alignments also seem to be of lower quality, in that they contain fewer informative sites and their alignments are less reliable (fig. 4*G–I*).

Finally, we tested for dependency between the taxonomical composition of the alignments and their classification as VGI or LGT and found these two properties are significantly dependent ($P < 0.001$, using χ^2 test). LGT alignments comprise about 30% of the archaeal or eubacterial alignments (as in the total data), but they are overrepresented in the universal group where they comprise 57% of the multiple alignments (fig. 4*J*). This leads us to conclude that the clustering of conserved sequences resulted in protein families that are not only large (as predicted by Elhaik et al. 2006) but also have a universal taxonomic distribution that covers much more diverse sequences and that seems to be the reason for their variability.

Our results so far suggest that alignments possessing properties that are known to increase the probability of obtaining incorrect branches are more frequent in the LGT group than in the VGI group. We then asked a slightly heretical question: Can we predict whether an alignment is likely to generate a tree with a strongly supported discordant branch on the basis of alignment properties alone? For this, we used an SVM classifier (Christiani and Shawe-Taylor 2000). In brief, a SVM is an algorithm that, provided with a learning set of features that might or might not correlate to a classificatory decision of the type "yes" or "no," gains experience with the learning set, and then is asked to classify objects, correctly if possible, on the basis of features alone. In the present case, the features correspond to alignment parameters as summarized in figure 1 and table 1, and the desired classification is the proper assortment of the alignment into LGT or VGI groups as predetermined by phylogenetic analysis. The classification performance is evaluated by its accuracy, sensitivity, and specificity (see Methods).

The SVM algorithm was thus trained and queried using the present alignments. In order to calculate SVM performance and standard deviations (SDs), we performed a 10-fold crossvalidation using 1/10 of the data in each step for training and the rest for testing. Accuracy, sensitivity, and specificity of the classifying process are to a vast extent influenced by the ratio of LGT/VGI in the training set. Accuracy and sensitivity are maximal when the proportion of LGT in the training set is equal to the total data (25%) and they decrease when higher LGT proportions are used. The specificity of the SVM classification is minimal at 25% LGT alignments in the training set and increases when higher proportions are used. When LGT proportion in the training set is fixed to 50%, all SVM performance measures are found in equilibrium (fig. 5*A*). The ratio of LGT/VGI alignments in the test set has no influence on the performance of the SVM classifier (fig. 5*B*). In our SVM classification procedure, we used training sets having an LGT/VGI ratio = 1 (see Methods). Performance of the classifying process was evaluated by trying all possible 2,047 combinations of the 11 properties to explore if there is a set of features that, if omitted from the training process, will deteriorate the results, or if there are some features that tend to impair the performance when included in the analysis.

Table 2 shows the combination of features that yielded the top performance values of accuracy, sensitivity, and specificity. We cannot really decide which is the best combination of feature vectors to be included in the training process because widely different combinations of features induce consistent results in the classification performance. But it seems that for equally high performance values for

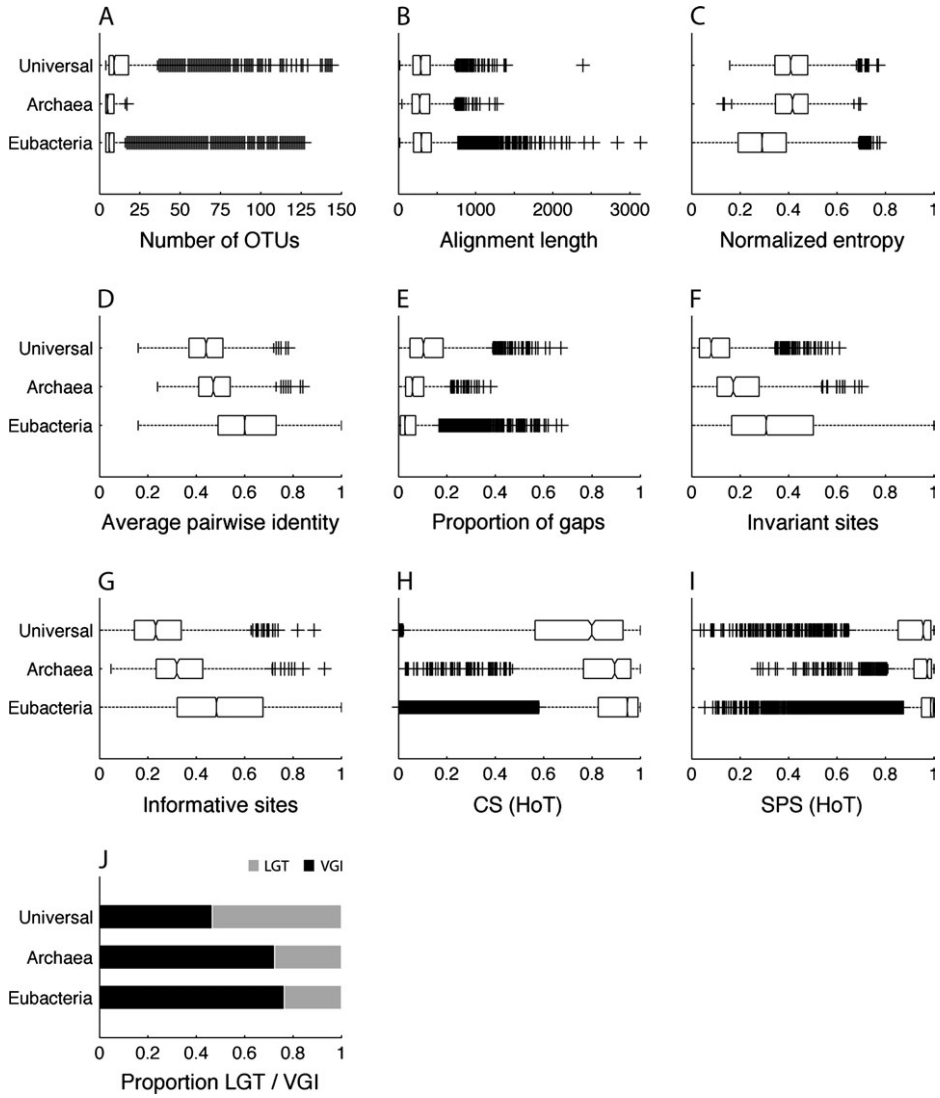


FIG. 4.—Differences in alignment properties for alignments containing only eubacterial sequences, only archaebacterial sequences, or sequences of both kingdoms (universal).

the three parameters (e.g., combinations yielding accuracy = 0.797 or accuracy = 0.796), the number of OTUs, entropy, average pairwise identity, and number of phyla are of particular importance. A complete table with all 2,047 tested combinations of features can be found in supplementary table S1, Supplementary Material online.

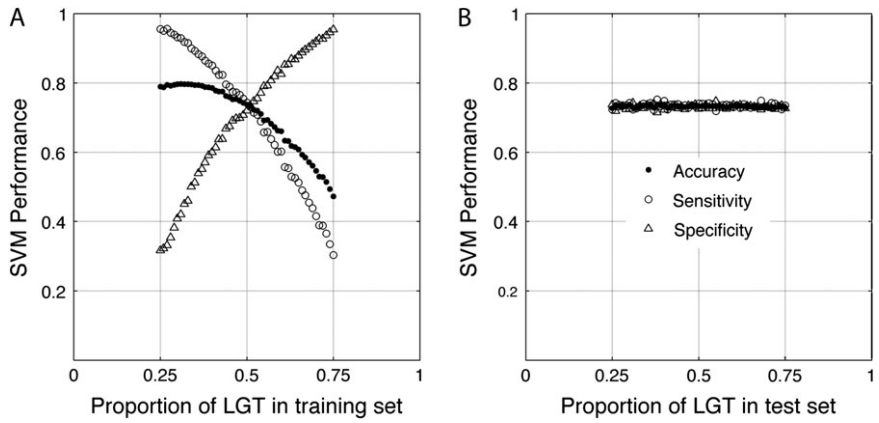


FIG. 5.—Performance of the classifier under different LGT proportions in the training set (*A*) and in the test set (*B*). In (*B*), the LGT/VGI ratio was adjusted to 1.

Table 2
Prediction of LGT/VGI Using an SVM Classifier Trained with Alignment Properties

Combination of Training Parameters											Performance		
Number of OTUs	Shannon Entropy	Average Pairwise Identity	Proportion of Gaps	Proportion of Invariant Sites	Proportion of Informative Sites	Alignment Length	CS (HoT)	SPS (HoT)	Number of Different Phyla	Kingdom Affiliation	Accuracy	Sensitivity	Specificity
v	*v*								*v*		0.797 ± 0.009	0.833 ± 0.020	0.692 ± 0.023
v	*v*							v	*v*		0.796 ± 0.009	0.835 ± 0.019	0.685 ± 0.021
v	*v*		*v*						*v*		0.796 ± 0.007	0.834 ± 0.015	0.687 ± 0.021
v	*v*				*v*				*v*		0.796 ± 0.009	0.830 ± 0.017	0.699 ± 0.017
v		*v*									0.794 ± 0.009	0.832 ± 0.020	0.688 ± 0.025
v		*v*			*v*				*v*		0.794 ± 0.012	0.828 ± 0.021	0.698 ± 0.018
v	*v*					*v*			*v*		0.794 ± 0.007	0.833 ± 0.015	0.685 ± 0.019
v	*v*	*v*			*v*				*v*		0.794 ± 0.012	0.827 ± 0.025	0.700 ± 0.025
v					*v*				*v*		0.794 ± 0.009	0.831 ± 0.019	0.687 ± 0.021
v		*v*					*v*		*v*		0.793 ± 0.008	0.829 ± 0.016	0.692 ± 0.020
										v	0.735 ± 0.009	0.931 ± 0.020	0.174 ± 0.022
							v	v	*v*	v	0.743 ± 0.008	0.887 ± 0.087	0.331 ± 0.247
								v		v	0.711 ± 0.041	0.884 ± 0.081	0.218 ± 0.074
			v					v	*v*	v	0.733 ± 0.012	0.873 ± 0.091	0.334 ± 0.229
v							*v*	v	*v*	v	0.765 ± 0.019	0.867 ± 0.093	0.474 ± 0.327
v							*v*	v	*v*		0.784 ± 0.018	0.851 ± 0.061	0.592 ± 0.215
v			*v*				*v*	v			0.771 ± 0.013	0.849 ± 0.066	0.551 ± 0.216
v			*v*					v	*v*		0.778 ± 0.015	0.848 ± 0.063	0.576 ± 0.224
			v				*v*	v	*v*	v	0.736 ± 0.012	0.848 ± 0.091	0.419 ± 0.235
v			*v*				*v*			v	0.762 ± 0.014	0.847 ± 0.082	0.519 ± 0.274
		v									0.544 ± 0.015	0.450 ± 0.031	0.812 ± 0.033
		v			*v*			v			0.598 ± 0.013	0.529 ± 0.025	0.794 ± 0.024
		v						v			0.576 ± 0.011	0.501 ± 0.023	0.790 ± 0.025
		v			*v*	*v*	*v*				0.586 ± 0.013	0.516 ± 0.027	0.787 ± 0.028
				v	*v*	*v*					0.665 ± 0.019	0.623 ± 0.034	0.786 ± 0.025
		v			*v*			v		v	0.601 ± 0.018	0.536 ± 0.037	0.785 ± 0.037
		v			*v*		*v*				0.590 ± 0.012	0.522 ± 0.028	0.785 ± 0.034
		v	*v*		*v*		*v*				0.592 ± 0.019	0.525 ± 0.040	0.783 ± 0.043
		v			*v*						0.582 ± 0.011	0.511 ± 0.024	0.782 ± 0.027
		v					*v*				0.562 ± 0.014	0.484 ± 0.030	0.782 ± 0.029

NOTE.—Alignment properties included in the training process are marked with *v*. The LGT/VGI ratio in the training set was adjusted to 1. Only combinations yielding the best 10 performance values for accuracy, sensitivity, and specificity are shown, respectively. Definitions of accuracy, sensitivity, and specificity can be found in the text. A table presenting the performance of all possible combinations is presented in the Supplementary Material online.

In other words, the alignment properties of the LGT and VGI groups, although having strongly overlapping distributions for all parameters (fig. 1; table 1), are nonetheless sufficiently different in a consistent manner that we can correctly predict about 78% of the time whether a Bayesian phylogenetic inference will generate a branch from a given alignment that is sufficiently discordant to be scored as an LGT. On the strength of this finding and circumstance that for each alignment parameter the LGT alignments were always skewed toward values that are known from simulation studies to generate incorrect branches (Nei 1996), it is likely that reliable construction of phylogenetic trees is affected and incorrectly reconstructed branches may be a possible source of LGT inference. The correlations are consistent with the view (Landan and Graur 2007) that sequence sets problematic at the level of alignments are likely to be problematic at the level of phylogenetic inference as well.

In principle, one could use our trained SVM on other alignment data sets in order to predict which alignments will result in discordant branches comparing with a reference tree. However, one would still have to distinguish between discordant branches stemming from either genuine LGTs or phylogenetic reconstruction artifacts. The results presented here indicate that the latter are more frequent in problematic alignments; hence, alignment quality has high impact on evolutionary inference from phylogenetic trees. A similar observation was recently presented for phylogenetic inference of ancient LGTs during the endosymbiosis of plastids (Deusch et al. 2008). This indicates that it is important to monitor and assess alignment quality in large-scale phylogenetic analyses, particularly those implementing automated or semiautomated phylogeny pipelines.

Supplementary Material

Supplementary tables S1–S4 and supplementary figures S1–S6 (and additional supporting figures) are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Giddy Landan and David Bryant for discussions. This work was funded by the Deutsche Forschungsgemeinschaft (M.R., W.M.), and the German-Israeli Foundation for scientific research and development (T.D.). We would like to thank the central computing facility of Düsseldorf University.

Literature Cited

Beiko RG, Chan CX, Ragan MA. 2005. A word-oriented approach to alignment validation. Bioinformatics. 21:2230–2239.

Beiko RG, Harlow TJ, Ragan MA. 2005. Highways of gene sharing in prokaryotes. Proc Natl Acad Sci. 102:4332–14337.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Christiani N, Shawe-Taylor J. 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge (MA): Cambridge University Press.

Delsuc F, Phillips MJ, Penny D. 2003. Comment on "Hexapod origins: monophyletic or paraphyletic?". Science. 301:1482d.

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 25:748–761.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. 32:1792–1797.

Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol. 23:1–3.

Felsenstein J. 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266:418–427.

Ge F, Wang LS, Kim J. 2005. The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biol. 3:e316.

Gotoh O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol. 264:823–838.

Grasso C, Lee C. 2004. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 20:1546–1556.

Hillis DM. 1995. Approaches for assessing phylogenetic accuracy. Syst Biol. 44:3–16.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucl Acids Res. 30:3059–3066.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.

Landan G, Graur D. Forthcoming. 2008. Characterization of pairwise and multiple sequence alignment errors. Gene, doi: 10.1016/j.gene.2008.05.016.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. 19:1–7.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632–1635.

Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet. 30:371–403.

Nei M, Takezaki N, Sitnikova T. 1995. Assessing molecular phylogenies. Science. 267:253–254.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.

Penny D, Hendy MD, Steel M. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol Evol. 7:73–79.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl Acids Res. 33:501–504.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res. 22:4673–4680.

Valdar WSJ. 2002. Scoring residue conservation. Proteins. 48:227–241.

Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. Science. 319:473–476.

Dan Graur, Associate Editor

Accepted May 5, 2009

5.3 Genomes of stigonematalean cyanobacteria (Subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids

Tal Dagan[1,⋆], Mayo Roettger[2,⋆], Karina Stucken[2], Giddy Landan[1,2], Robin Koch[2], Peter Major[1], Sven B. Gould[1], Vadim V. Goremykin[3], Rosmarie Rippka[4], Nicole Tandeau de Marsac[4,10], Muriel Gugger[5], Peter J. Lockhart[6], John F. Allen[7,8], Iris Brune[9], Irena Maus[9], Alfred Pühler[9], and William F. Martin[1]

[1]Institut für Genomische Microbiologie, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

[2]Institut für Molekulare Evolution, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

[3]IASMA Research and Innovation Center, Fondazione Edmund Mach, San Michele all'Adige (TN), Italy

[4]Institut Pasteur, Unité des Cyanobactéries, Paris, France

[5]Institut Pasteur, Laboratoire Collection des Cyanobactéries, Paris, France

[6]Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

[7]School of Biological and Chemical Sciences, Queen Mary, University of London, London, United Kingdom

[8]Research Department of Genetics Evolution and Environment, University College London, London, United Kingdom

[9]Center for Biotechnology, University of Bielefeld, Bielefeld, Germany

[10]Present address: Aix-Marsaille University, Laboratoire de Chimie Bactérienne (LCB), Marsaille, France

[⋆]Der Beitrag beider Autoren ist gleichwertig.

Der vorgestellte Artikel wurde in der Fachzeitschrift *Genome Biology and Evolution* veröffentlicht (Dagan *et al.*, 2012).

Beitrag von Mayo Röttger, gleichberechtigter Erstautor:

Versuchsplanung:	30 %
Datenanalyse:	70 %
Verfassen des Manuskripts:	30 %

GENOME BIOLOGY AND EVOLUTION

SMBE

GBE

Genomes of Stigonematalean Cyanobacteria (Subsection V) and the Evolution of Oxygenic Photosynthesis from Prokaryotes to Plastids

Tal Dagan[1,*,†], Mayo Roettger[2,†], Karina Stucken[1], Giddy Landan[1,2], Robin Koch[1], Peter Major[2], Sven B. Gould[2], Vadim V. Goremykin[3], Rosmarie Rippka[4], Nicole Tandeau de Marsac[4,10], Muriel Gugger[5], Peter J. Lockhart[6], John F. Allen[7,8], Iris Brune[9], Irena Maus[9], Alfred Pühler[9], and William F. Martin[2]

[1]Institute of Genomic Microbiology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

[2]Institute of Molecular Evolution, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

[3]IASMA Research and Innovation Center, Fondazione Edmund Mach, San Michele all'Adige (TN), Italy

[4]Institut Pasteur, Unité des Cyanobactéries, Paris, France

[5]Institut Pasteur, Laboratoire Collection des Cyanobactéries, Paris, France

[6]Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

[7]School of Biological and Chemical Sciences, Queen Mary, University of London, London, United Kingdom

[8]Research Department of Genetics Evolution and Environment, University College London, London, United Kingdom

[9]Center for Biotechnology, University of Bielefeld, Bielefeld, Germany

[10]Present address: Aix-Marseille University, Laboratoire de Chimie Bactérienne (LCB), Marseille, France

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: tal.dagan@hhu.de; tal.dagan@uni-duesseldorf.de.

Accepted: December 4, 2012

Data deposition: Genomes have been deposited in NCBI under accessions PRJNA104961, PRJNA104963, PRJNA104969, PRJNA104967, PRJNA104965, and PRJNA157363.

Abstract

Cyanobacteria forged two major evolutionary transitions with the invention of oxygenic photosynthesis and the bestowal of photosynthetic lifestyle upon eukaryotes through endosymbiosis. Information germane to understanding those transitions is imprinted in cyanobacterial genomes, but deciphering it is complicated by lateral gene transfer (LGT). Here, we report genome sequences for the morphologically most complex true-branching cyanobacteria, and for *Scytonema hofmanni* PCC 7110, which with 12,356 proteins is the most gene-rich prokaryote currently known. We investigated components of cyanobacterial evolution that have been vertically inherited, horizontally transferred, and donated to eukaryotes at plastid origin. The vertical component indicates a freshwater origin for water-splitting photosynthesis. Networks of the horizontal component reveal that 60% of cyanobacterial gene families have been affected by LGT. Plant nuclear genes acquired from cyanobacteria define a lower bound frequency of 611 multigene families that, in turn, specify diazotrophic cyanobacterial lineages as having a gene collection most similar to that possessed by the plastid ancestor.

Key words: plastid evolution, endosymbiosis, phylogenomics, true-branching cyanobacteria, nitrogen fixation.

Introduction

Cyanobacteria are crucial players in Earth and life history because they generated the oxygen that has been present in the Earth's atmosphere for the last 2.4 billion years (Bekker et al. 2004) and because one uniquely fateful cyanobacterium became, via endosymbiosis, the ancestor of all plastids among photosynthetic eukaryotes (Gould et al. 2008). Though they continue to impact global geochemical cycles through N_2-fixation (Moisander et al. 2010), and the

sequestering of trace metals (Morel and Price 2003) as well as phosphorous (van Mooy et al. 2009), their main ecological significance is the oxygen-producing photosynthetic apparatus that fuels most contemporary food chains. Their main evolutionary significance is that they mediated two pivotal innovations in life's history—water-splitting photosynthesis and the origin of primary plastids. Clues to both of those major evolutionary transitions should, in principle, be imprinted in cyanobacterial genomes. But reconstructing those events is not straightforward, because lateral gene transfer (LGT) redistributes genes among prokaryote genomes (Ochman et al. 2000), and among cyanobacterial genomes in particular (Raymond et al. 2002; Mulkidjanian et al. 2006; Dufresne et al. 2008; Shi and Falkowski 2008), over geological time.

By necessity, and perhaps more so than for any other prokaryotic group, LGT has always been hard-wired into the bigger picture of cyanobacterial evolution. To explain the origin of cyanobacterial water-splitting photosynthesis, both of the main competing theories require LGT to account for the distribution of photosystems across prokaryotic groups (Xiong and Bauer 2002; Hohmann-Marriot and Blankenship 2011). This is because the reaction centers of photosystems I and II clearly share common ancestry (Baymann et al. 2001; Hohmann-Marriot and Blankenship 2011), but without specifying how they entered the cyanobacterial ancestor genome. One theory posits that the two photosystems evolved in independent lineages and became merged in the founder cyanobacterium via LGT (Baymann et al. 2001), while the alternative has it that the photosystems diverged within a photosynthetic (protocyanobacterial) ancestor and were subsequently exported via LGT to some anoxygenic photosynthetic lineages (Xiong and Bauer 2002; Mulkidjanian et al. 2006; Sharon et al. 2009). Compatible with a role for LGT in photosystem evolution is the finding that the genes for both photosystems I and II are mobile in marine phage metagenomes (Lindell et al. 2004; Sharon et al. 2009).

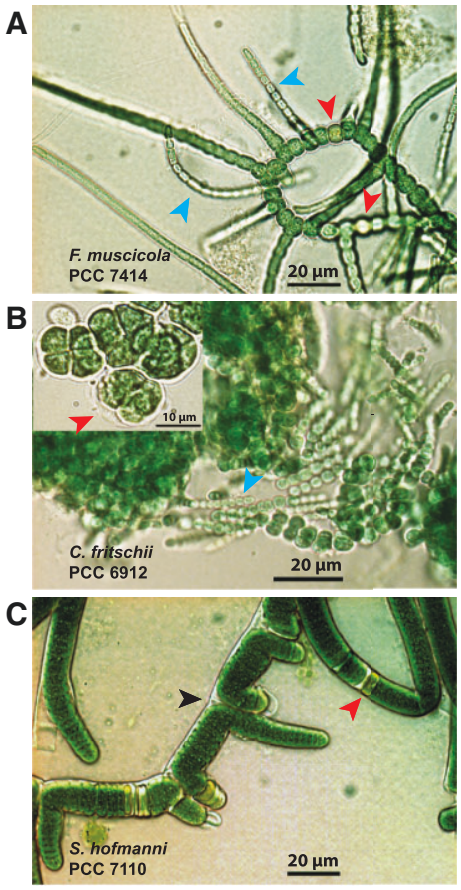
LGT also figures into the origin of plastids, because many genes were transferred from endosymbiont to host. Chloroplasts were once free-living cyanobacteria and contained approximately 2,000 proteins (Richly and Leister 2004), a number comparable with a cyanobacterium, yet the genomes of modern plastids contain only 5–10% as many genes as those of their free-living cousins. This suggests that hundreds or thousands of the plastid ancestor's genes were either lost or relocated to the host nucleus during the course of plant evolution via endosymbiotic gene transfer (EGT) (Gould et al. 2008). Furthermore, the phylogenetic identity of the plastid ancestor remains debated because of LGT. Different phylogenetic trees trace the plastid ancestor near the base of cyanobacterial diversification (Criscuolo and Gribaldo 2011), near coccoid cyanobacteria within the *Synechococcus–Prochlorococcus* (SynPro) clade (Reyes-Prieto et al. 2010), near the nitrogen-fixing *Cyanothece* clade (Deschamps et al. 2008), or near filamentous, heterocyst-forming cyanobacterial lineages (Deusch et al. 2008). The simplest explanation for such findings—in an evolutionary context that incorporates LGT—is that the plastid ancestor donated one (chimeric) genome's worth of genes to the host, and that LGT has been reassorting the homologs of these genes among free-living cyanobacterial and other prokaryote genomes ever since (Deusch et al. 2008). Because of LGT over time, the question of which "lineage" of cyanobacteria gave rise to the plastid loses meaning (Doolittle and Bapteste 2007), because the genomes and nature of the "lineages" have changed since the time of plastid origin over 1.2 billion years ago (Deusch et al. 2008; Gross et al. 2008). However, comparison of plant genes acquired from the plastid ancestor with cyanobacterial homologs can reveal which modern cyanobacteria harbor a collection of genes most similar to that of the plastid ancestor.

So far, missing in genomic studies of cyanobacterial evolution are sequences from the group designated as subsection V (Rippka et al. 1979). Subsection V cyanobacteria grow as filaments that differentiate heterocysts (specialized N_2-fixing cells), they produce cyst-like resting cells (akinetes) as well as differentiated motile trichomes (hormogonia), and most exhibit true branching. The developmental and morphological variety of subsection V cyanobacteria places them among the most complex of prokaryotes, for which reason they were even long thought to be the direct ancestors of all eukaryotes but only in the days before the endosymbiotic origin of plastids has been postulated (Mereschkowsky 1905) and eventually gained compelling support (Doolittle 1980). To better understand the role of subsection V species in cyanobacterial evolution and their possible relationship to the plastid ancestor, we have sequenced five genomes sampling a broad spectrum of filamentous, true-branching architecture (fig. 1*A* and *B*), and diverse geographical locations including rice fields in India (*Fischerella muscicola* PCC 73103 and *Chlorogloeopsis fritschii* PCC 6912), and hot springs in New Zealand (*F. muscicola* PCC 7414), Wyoming, USA (*F. thermalis* PCC 7521), and in Spain (*C. fritschii* PCC 9212) (Rippka et al. 1979). In addition *Scytonema hofmanni* PCC 7110, a Nostocales representative (subsection IV) isolated from a limestone cave (Crystal cave, Bermuda) (Rippka et al. 1979), whose filaments form false branches (fig. 1*C*) and exhibit aerial growth, was included for comparison.

Materials and Methods

Cyanobacterial Cultures and DNA Isolation

Stock cultures were maintained at 37°C on slants (or plates) in BG11o medium (Rippka and Herdman 2002), supplemented with 5 mM $NaHCO_3$ and solidified with 0.9% (w/v) washed agar (Sigma, A 8678). For DNA isolation, cultures were grown at 37°C in BG11 medium (Rippka and Herdman 2002), with orbital shaking (100 rpm) in an Infors Incubator, at a PPFD of



D

Genus	*Chlorogloeopsis*		*Fischerella*			*Scytonema*
Species	*fritschii*	*fritschii*	*muscicola*	*muscicola*	*thermalis*	*hofmanni*
PCC strain	6912	9212	73103	7414	7521	7110
Assembled scaffolds	8	21	37	16	4	81
Bases in scaffolds	7,819,561	7,698,501	7,573,287	7,008,859	5,482,660	12,073,012
%GC	41.5	41.5	38.6	40.4	40.5	41.5
No. CDSs	7,439	7,571	7,231	7,167	5,340	12,356
CDS density (%)	82	82	78	79	79	83
Mean gene length (bp)	830	829	820	778	825	806
rRNA operons	4	3	3	3	4	3
No. tRNA	45	41	43	59	42	67
No. Plasmids	2	n. d.[(1)]	n. d.[(1)]	0	0	n. d.[(1)]

[(1)] Not determined.

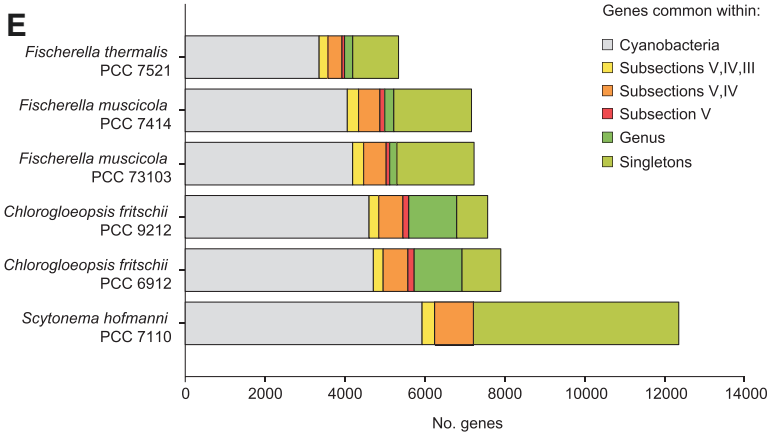


Fig. 1.—Genomes of Stigonematales and *Scytonema*. (*A*) *Fischerella muscicola* PCC 7414, forming true lateral branches. (*B*) *Chlorogloeopsis fritschii* PCC 6912, undergoing cell divisions in more than one plane but never producing lateral branches. Heterocysts and hormogonia, differentiated by members of both genera are marked by red and cyan arrows, respectively. (*C*) *Scytonema hofmanni* PCC 7110 showing false branching filaments (black arrow) and heterocysts (red arrow). (*D*) Genomic features of the six novel sequenced genomes. Genomes have been deposited in NCBI under accessions (PRJNA104961, PRJNA104963, PRJNA104969, PRJNA104967, PRJNA104965, and PRJNA157363). Fully annotated versions are available at www.molevol.de/resources. (*E*) Frequency distribution of protein coding genes in the new genomes, and comparison with other cyanobacterial genomes examined.

30 µmol quanta $m^{-2}s^{-1}$. Cultures were harvested after 3–6 weeks of incubation, depending on density of the inoculum and the growth rates of the strains. DNA isolation from strains of *Chlorogloeopsis* was performed as described (Franche and Damerval 1988), with the addition of 1% Sarkosyl during lysozyme treatment to remove polysaccharides and a final RNA digestion step. Polysaccharide-free high molecular weight genomic DNA (gDNA) from strains of *Fischerella* was obtained by following a protocol for polysaccharide-rich plants (Sharma et al. 2002).

Genome Sequencing and Annotation

Prior to genome sequencing the identity of the gDNA was verified by sequencing of the 16S rDNA with primers 101F (ACTGGCGGACGGGTGAGTAA) and 1047R (GACGACAGCC ATGCAGCACC), and comparison against cyanobacterial sequences available in NCBI. Genome sequencing was performed on the Genome Sequencer FLX using Titanium chemistry (Roche Applied Science, Penzberg, Germany) yielding a 10- to 32-fold coverage. Genome scaffolding was achieved by 3 kbp paired-end standard runs. The sequencing libraries were prepared from 4 µg of gDNA for whole genome shot gun sequencing and 5 µg of gDNA for paired-end sequencing, according to the supplier's instructions. Additionally, a fosmid library was constructed with the Copy Control Fosmid Library Production Kit (Biozym Scientific, Hess. Oldendorf, Germany). Terminal DNA sequences of cloned genomic inserts were determined with an ABI 3730xl DNA Analyzer (Life Technologies, Darmstadt, Germany). Furthermore, Sanger-reads were generated from fosmid clones to cover the gaps between contigs for each of the five genomes. Sequence data were assembled with the GS De Novo Assembler Software (ver. 2.0.01.14, 2.3, and 2.5.3). For each genome, large (>500 bp) and small contigs

(<500 bp) were obtained, including numerous repetitive elements and insertion segments. For finishing purposes, all DNA sequences were uploaded into the Consed program (Gordon et al. 1998). The final annotation including COGs (Tatusov et al. 2001) of the genome sequences was accomplished with the GenDB software (Meyer et al. 2003). Gene prediction was performed by means of combining results of the software tools GLIMMER (Delcher et al. 1999), CRITICA (Badger and Olson 1999), and GISMO (Krause et al. 2007).

Phylogenetic Analysis of Cyanobacterial Genomes

Fully sequenced cyanobacterial proteomes were downloaded from NCBI version March/2011. For the reconstruction of cyanobacterial gene families, we conducted an all-against-all BLAST search (Ver. 2.2.17) (Altschul et al. 1997) using the protein sequences. Reciprocal best BLAST hits (rBBH) were performed using a threshold of E value $\leq 10^{-10}$ and percent amino acid identity $\geq$30. For the clustering analysis, the overall protein sequence similarity between rBBH proteins, calculated as the percent of identical amino acids, was multiplied by the length ratio of the two proteins. Clusters of gene families were inferred from the rBBH similarity matrix using the MCL ver. 1.008 clustering procedure (Enright et al. 2002), with the inflation parameter (I) set to 2.0. For the reconstruction of a consensus tree phylogeny, 324 gene families present as single copies in all cyanobacterial genomes analyzed were aligned with MAFFT (Katoh et al. 2002) ver. 6.717b. Phylogenetic trees were reconstructed using the Neighbor-Joining (NJ) approach (Saitou and Nei 1987). Protein sequence distances were calculated with PROTDIST (Felsenstein 1993), and applying the JTT substitution model (Jones et al. 1992). Phylogenetic trees were reconstructed with NEIGHBOR (Felsenstein 1993). The consensus phylogeny was reconstructed with CONSENSE (Felsenstein 1993). A concatenated alignment was reconstructed from the aligned protein sequences, and all genes were weighted equally (supplementary fig. S1, Supplementary Material online). A phylogenetic tree was reconstructed from the concatenated alignment using the NJ approach and the software described as earlier. A phylogenetic network was reconstructed with SplitsTree Ver. 4.10 using the default parameters (Huson and Bryant 2006). A minimal lateral network (MLN) was reconstructed using the consensus phylogeny as the reference tree, and the gene families described earlier according to the approach described in Dagan et al. (2008). Maximum likelihood phylogeny was reconstructed using PhyML (Guindon et al. 2010) with LG model + I (estimation of invariant sites) + G (gamma distribution with 4 rate categories). Tree topology (SPR), branch length, and rate parameters were optimized.

Phylogenetic Analysis of the Plastid Ancestor

Sequences of nuclear-encoded proteins from the whole genomes of *Arabidopsis thaliana, Oryza sativa* subsp. *japonica*, *Physcomitrella patens, Chlamydomonas reinhardtii, Entamoeba histolytica, Dictyostelium discoideum, Filobasidiella neoformans, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster, Ciona intestinalis, Danio rerio, Gallus gallus, Canis lupus familiaris*, and *Homo sapiens* were obtained from RefSeq database release November 2009 (Pruitt et al. 2007). Nuclear proteomes of *Cyanidioschyzon merolae* version February 2005 (Matsuzaki et al. 2004), *Ostreococcus tauri* version 2.0 (Palenik et al. 2007), and *Xenopus tropicalis* release 4.1, August 2005 (Bowes et al. 2008), were downloaded from the respective genome project websites. Additionally, 650 fully sequenced genomes of prokaryotes, including those of 46 cyanobacterial representatives, were downloaded from NCBI RefSeq database release November 2009 (Pruitt et al. 2007). To avoid clustering artifacts of distantly related eukaryotic and prokaryotic sequences, the sequences of cyanobacteria and photosynthetic eukaryotes were first clustered into separate sets of protein families. Matrices of algal/plant and cyanobacterial sequences were constructed from reciprocal best BLAST hits using an all-against-all BLAST, and thresholds of E-value $\leq 10^{-10}$ and amino acid sequence identities $\geq$25%. Clusters of homologous protein sequences were reconstructed from each of the matrices using MCL (Enright et al. 2002) Ver. 08-312, 1.008, with scheme = 7 and $I = 2.0$. Protein sequences of noncyanobacterial prokaryotes and nonphotosynthetic eukaryotes were added to the plant/algal clusters of proteins, depending on their sequence homologies using the above threshold, and a limit of three sequences per phylum. Overlapping plant/algal and cyanobacterial clusters were joined. The sequences of protein families were aligned using MAFFT (Katoh et al. 2002) Ver. 6.717b (2009/12/03). Multiple sequence alignment quality was assessed using the HoT-method (Landan and Graur 2007). Plant/algal protein sequences with Sum of Pairs Score <80% were excluded from the cluster. Phylogenetic trees were reconstructed using maximum likelihood approach with PhyML (Guindon et al. 2010) and the best-fit model as inferred with ProtTest (Abascal et al. 2005). The search for a best-fit model using ProtTest was restricted for nuclear gene substitution models including JTT (Jones et al. 1992) and WAG (Whelan and Goldman 2001) matrices. These were tested with all combinations of +I (estimation of invariant sites), +G (gamma distribution with 4 rate categories), and +F (using amino acid frequencies from the alignment) parameters. Branch lengths, model, and topology were optimized. From among 35,862 trees in total, WAG model was found as the best fit in 89% of the trees, with WAG + I + G as the more prevalent choice (34%). Genes of endosymbiotic origin in algal and plant genomes were inferred from the phylogenetic trees by searching for sisterhood between cyanobacterial protein sequences and their counterparts encoded by the nuclear genes of the photosynthetic eukaryotes (Martin et al. 2002). Protein families in the latter phototrophs were counted as having resulted from EGT(s), if

at least one of them had a cyanobacterial sequence as the nearest neighbor. Concatenated alignments were analyzed and used for tree construction by the same methods as described earlier.

Results

Genomes of Subsection V (Stigonematales) and *Scytonema*

The genome size distribution of the five Stigonematales strains (5.9 ± 2 Mb; fig. 1*D*) is similar to that of subsection IV members (Nostocales) (Larsson et al. 2011). With only 5,340 CDSs, *F. thermalis* PCC 7521 has the smallest genome among the subsection V members, whereas the genome of *S. hofmanni* PCC 7110 (subsection IV) has 12,356 predicted ORFs, making it the most gene-rich prokaryote sequenced to date (fig. 1*D*). Clustering of all 223,941 CDSs encoded in 51 cyanobacterial genomes by protein sequence similarity resulted in 18,185 cyanobacterial protein families and 47,174 singletons. Protein families with metabolic or cellular functions have significantly more duplicates in strains of subsection V than in those of subsection IV ($P < 2.2 \times 10^{-16}$, paired *t* test). Subsection V and IV strains do not differ in gene copy number for information processing protein families ($P = 0.11$, paired *t* test). The genome of strain PCC 7521 contains fewer duplicates ($P < 2.2 \times 10^{-16}$, paired *t* test) than the other two representatives of *Fischerella*. The frequency of genes shared with other filamentous cyanobacteria and the distribution of gene function are similar (fig. 1*E* and supplementary fig. S2, Supplementary Material online) among the three phenotypically similar *Fischerella* strains (Rippka et al. 1979).

Patterns of gene presence and absence might identify genes related to cyanobacterial morphological diversity (Stucken et al. 2010; Larsson et al. 2011). A subset of 22 protein families is unique and common to all filamentous cyanobacteria in our sample (supplementary table S1, Supplementary Material online), only few of which have known function. Subsection V members share 7 ± 1% of their proteome with those of subsection IV, and 73 protein families are specific to heterocyst-forming strains (supplementary table S2, Supplementary Material online). Most of the remaining subsections IV- and V-specific genes fall into cell wall, membrane, and envelope biogenesis COGs, such as glycosyltransferases, exopolysaccharide synthesis, and secretion. Some of the subsection V-specific protein families might be involved in the multiseriate filament phenotype and formation of true branches. On average, only 2% of the proteins encoded in subsection V genomes are specific to true-branching forms. Only 46 gene families are uniquely shared among subsection V genomes (supplementary table S3, Supplementary Material online). Although their functions are yet unknown, their classifications entail mostly cell wall, membrane, envelope biogenesis, and signal transduction functions. The relative paucity of proteins comprising the core set of the true branching cyanobacteria suggests that this phenotype hinges upon very few expressed proteins, which may mainly affect regulation of cell division genes and/or localization of their products.

Vertical and Lateral Components of Cyanobacterial Genome Evolution

To reconstruct a cyanobacterial backbone phylogeny, we identified all 324 single-copy protein families common to all 51 cyanobacteria in our sample and reconstructed their phylogenetic trees. The consensus tree (fig. 2), rooted with *Gloeobacter violaceus*, indicates a single origin for the filamentous architecture, and the concatenated alignment (564,408 sites) yielded an identical topology with NJ (supplementary fig. S3, Supplementary Material online), where all branches are supported by 100% bootstrap replicates. Maximum likelihood reconstruction yielded a phylogeny in which filamentous cyanobacteria are polyphyletic (supplementary fig. S3, Supplementary Material online), the difference to NJ being the position of *Microcoleus chthonoplastes* PCC 7420, a filamentous strain isolated from salt marshes (Rippka et al. 1979). Current whole-genome cyanobacterial phylogenies group *Microcoleus* with subsection I (Criscuolo and Gribaldo 2011), yielding paraphyly for filamentous forms. Although 55 of our 324 single copy gene trees support that position for *Microcoleus*, 111 recover filamentous monophyly, discrepancies that might reflect the workings of LGT (Raymond et al. 2002; Mulkidjanian et al. 2006; Shi and Falkowski 2008; Dufresne et al. 2008). To test the consistency of the backbone (consensus) phylogeny, we reconstructed a phylogenetic network using SplitsTree (Huson and Bryant 2006). The resulting network reveals a paucity of conflicting splits in the data (supplementary fig. S4, Supplementary Material online). A total of 92 out of 212 splits are compatible with the NJ tree topology and their sum of split weight amounts to 96% of the total network; and thus, the NJ tree explains most of the split variability in the data.

To estimate the degree and distribution of LGT in cyanobacterial evolution, we reconstructed a MLN, which infers LGT frequencies by allowing increasing amounts of LGT per protein family across a given backbone phylogeny (here the consensus tree), and identifying for all gene families the LGT frequency at which the distributions of modern genome sizes and inferred ancestral genome sizes agree best (Dagan et al. 2008). The MLN analysis conservatively assumes that all gene trees for all protein families are compatible (Dagan et al. 2008) and entails no gene tree comparisons. It revealed that 6,068 (34%) of the cyanobacterial protein families require no LGT to account for their gene distributions, whereas 12,116 (66%) protein families have undergone at least one LGT event. Because the method does not tally conflicting gene trees for

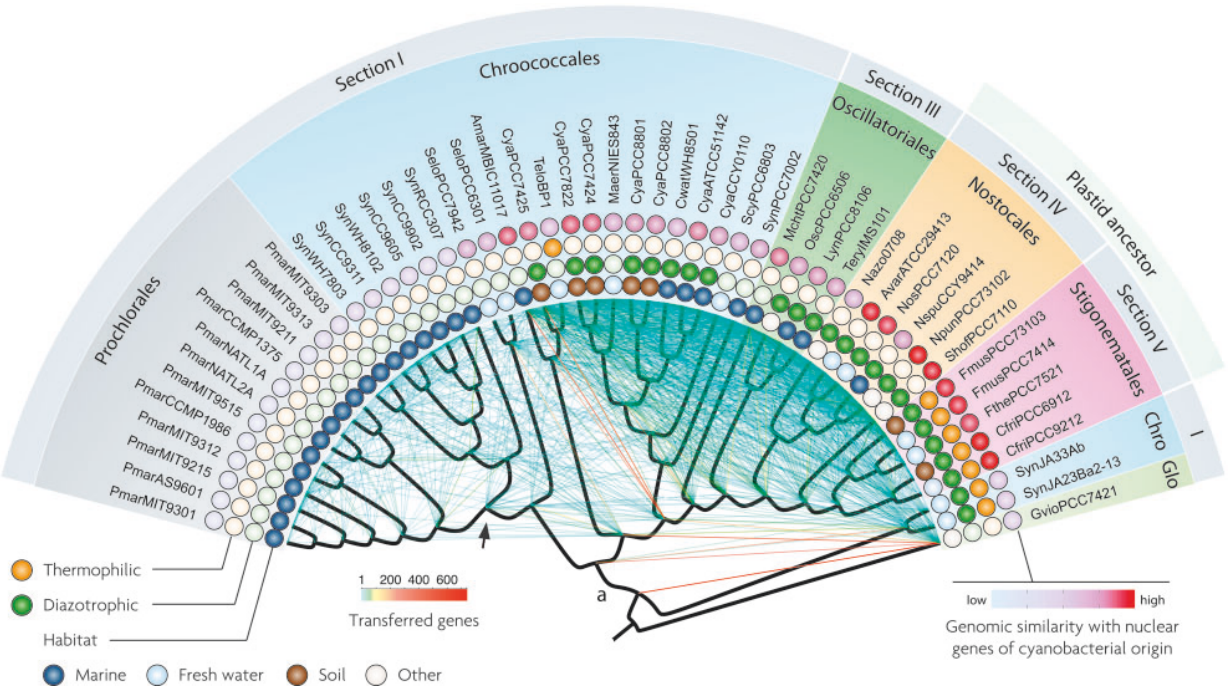


Fig. 2.—Vertical and lateral gene evolution in cyanobacterial genomes. NJ consensus (or backbone) tree, inferred from 324 single-copy protein families common to all 51 cyanobacteria in our sample, and rooted with *Gloeobacter violaceus* PCC 7421. Branches indicating vertical gene evolution are indicated in black. The MLN is indicated by edges that do not map onto the vertical component, with number of genes per edge indicated by a color gradient from cyan (1 gene) to orange (736 genes). The phylogenetic position of the eukaryotic clade reconstructed using 23 core genes is marked by "a." The SynPro clade is marked by an arrow.

homologous sequences, these are conservative lower bound estimates, in contrast to other recent studies (Raymond et al. 2002; Mulkidjanian et al. 2006; Shi and Falkowski 2008; Dufresne et al. 2008). Our estimate is found in agreement with earlier quantification of LGT frequency among cyanobacteria using an embedded quartets approach (Zhaxybayeva et al. 2006).

The MLN is presented in figure 2, and shows vertical components of cyanobacterial evolution and a network of 1,183 edges indicating laterally shared genes. Within the network, 358 edges (32%) represent a single laterally shared gene, whereas most edges (55%) carry $\leq$3 genes. Only 91 (7%) of edges carry >20 genes. Thus, bulk transfers of tens of genes or more are rare. The clade of marine *Prochlorococcus* and *Synechococcus* (SynPro) strains, which are recognized as being closely related environmental specialists of reduced genome size (Rocap et al. 2003; Dufresne et al. 2008), appear to have the lowest LGT frequency. The intertwined phylogenies within this clade (Zhaxybayeva et al. 2009) go undetected because the MLN is reconstructed from gene presence/absence data that are uninformative for the reconstruction of recombination events at the intra-species level (Dagan et al. 2008). The most highly connected nodes implicate the four contemporary strains *Acaryochloris marina* MBIC 11017, *Cyanothece* PCC 7425, *M. chthonoplastes* PCC 7420, and *S. hofmanni* PCC 7110 (fig. 2). Two of these strains, *A. marina*, an atypical marine unicellular cyanobacterium producing chlorophyll *d* as the primary photosynthetic pigment (Swingley et al. 2008), and *M. chthonoplastes*, a marine mat former, have the largest genomes (8.36 and 8.65 Mb, respectively) known for members of subsections I and III, and show an expansion of protein families (Larsson et al. 2011). The MLN pinpoints large genomes as harboring gene pools that are frequently transferred among cyanobacteria, and identifies subsection V strains as being more highly connected with strains of subsections IV and III (1.4 edges/node) than with unicellular strains (0.3 edges/node), also when strains of the SynPro clade are excluded (0.7 edges/node). This may suggest the existence of a LGT barrier between unicellular (mostly marine) and filamentous (mostly terrestrial) cyanobacteria.

The Nature of the Plastid Ancestor

To identify plant nuclear genes of cyanobacterial origin, we reconstructed 35,862 phylogenetic trees containing both eukaryotic and prokaryotic homologs and looked for trees in which plants and cyanobacteria branch together. In the present sample, considering all trees, between 8.7% and 11.5% of all nuclear genes in photosynthetic eukaryotes sampled branch with cyanobacterial homologs (table 1).

Table 1
Proportion of Plant Genes of Endosymbiotic Origin

	No. Proteins	Total Tree Set			CS ≥ 80%		≤ 3 homologues	
		No. Trees	No. Putative EGT	EGT Bootstrap Support	No. Trees	No. Putative EGT	No. Protein Families	No. Putative EGT
Arabidopsis thaliana	30,897	9,025	801 (8.9%)	87.89 ± 20.10	3,306	424 (12.8%)	2,091	136 (6.5%)
Oryza sativa	26,712	7,292	637 (8.7%)	84.82 ± 21.41	2,596	347 (13.4%)	1,623	95 (5.9%)
Physcomitrella patens	35,468	8,847	903 (10.2%)	84.74 ± 22.11	3,425	542 (15.8%)	1,402	78 (5.6%)
Ostreococcus tauri	7,715	3,495	403 (11.5%)	84.64 ± 21.20	1,232	247 (20.1%)	324	26 (8.0%)
Cyanidioschyzon merolae	4,761	2,688	307 (11.4%)	83.92 ± 20.05	844	167 (19.8%)	223	15 (6.7%)
Chlamydomonas reinhardtii	14,262	4,515	478 (10.6%)	83.81 ± 21.04	1,646	283 (17.2%)	599	41 (6.8%)
Total	119,815	35,862	3,529 (9.8%)	84.97 ± 20.98	13,049	2,010 (15.4%)	6,262	391 (6.2%)

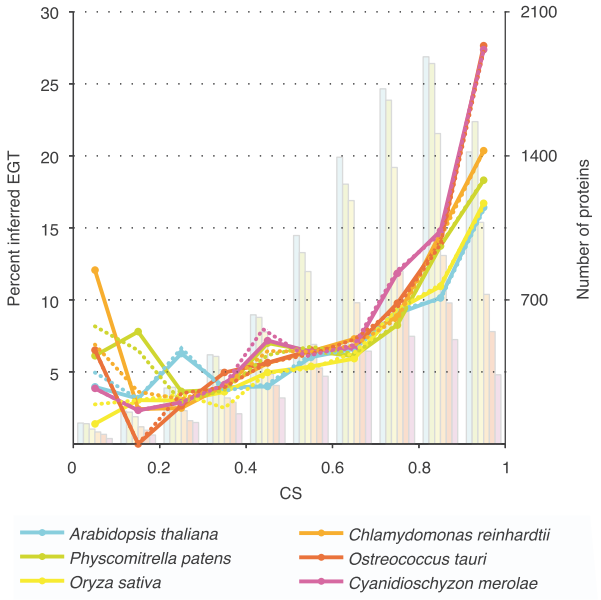


Fig. 3.—Phylogenetic characteristics of EGT inference. The frequency of EGT as inferred from alignments of varying reliability degrees. The distribution of alignment reliability as estimated by column score (CS) is presented in bars, colored according to the respective eukaryotes. The CS measure is calculated as the proportion of alignment sites whose reconstruction is independent upon the direction upon which the sequences are fed to the alignment algorithm (Landan and Graur 2007). The frequency of genes inferred as EGT is plotted above in the eukaryote-dependent strongly colored lines, with the proportions inferred from trees reconstructed by maximum likelihood and NJ approaches in solid and dashed lines, respectively.

For the most reliable alignments, where false negatives are less likely, the proportion of genes acquired from plastids ranges between 16% of the genes in *Arabidopsis* genome and >20% of the genes in the smaller genomes of *Ostreococcus* and *Cyanidioschyzon* (fig. 3), with energy metabolism and carbohydrate metabolism (99 genes) being the most frequent functional categories (supplementary fig. S5, Supplementary Material online). Clearly, the quantitative contribution of cyanobacteria to plant genomes was great, and the backbone of plant metabolism was acquired from them—plants are, biochemically, cyanobacteria wrapped in a bigger box.

To trace the nature of the plastid ancestor, we first assembled a dataset of 23 nuclear genes of plastid origin present in all plant and cyanobacterial genomes sampled. The tree of concatenated alignments, rooted by *G. violaceus* PCC 7421, shows a deep branch placing plastids basal among cyanobacteria (designated with an "a" in fig. 2). Expanding the data set to include 200 universal cyanobacterial gene families with a single, composite plant OTU (genes acquired from cyanobacteria and present in at least one plant) yielded the same long, deep branch. Long basal branches are characteristic of long-branch attraction (LBA), a well-known phylogenetic artifact. Compositional heterogeneity such as AT bias and heterotachy can cause LBA (Lockhart et al. 2006), and a basal position due to an LBA often involves the grouping of strains in which strain-specific character states are abundant (Stiller and Hall 1999). The sequences of the 23 universally distributed proteins in the six photosynthetic eukaryotes were found to contain significantly more unique substitutions than their cyanobacterial homologues ($P = 7 \times 10^{-66}$, one-tailed Kolmogorov–Smirnov test, supplementary fig. S6, Supplementary Material online), and an examination of the larger set of 200 phylogenetic trees reconstructed for genes of endosymbiotic origin shows that the eukaryotic clade branch length is on average 10-fold larger than that of the cyanobacterial branches. The basal position of plastids among cyanobacteria in the concatenated alignment tree (fig. 2 and supplementary fig. S6, Supplementary Material online) is attributable to LBA. Worse, given that LGT is frequent among cyanobacteria (Raymond et al. 2002; Mulkidjanian et al. 2006; Shi and Falkowski 2008; Dufresne et al. 2008), there is no reason to suspect that any "core" gene phylogeny will be a faithful proxy for the rest of the genome (Doolittle and Bapteste 2007).

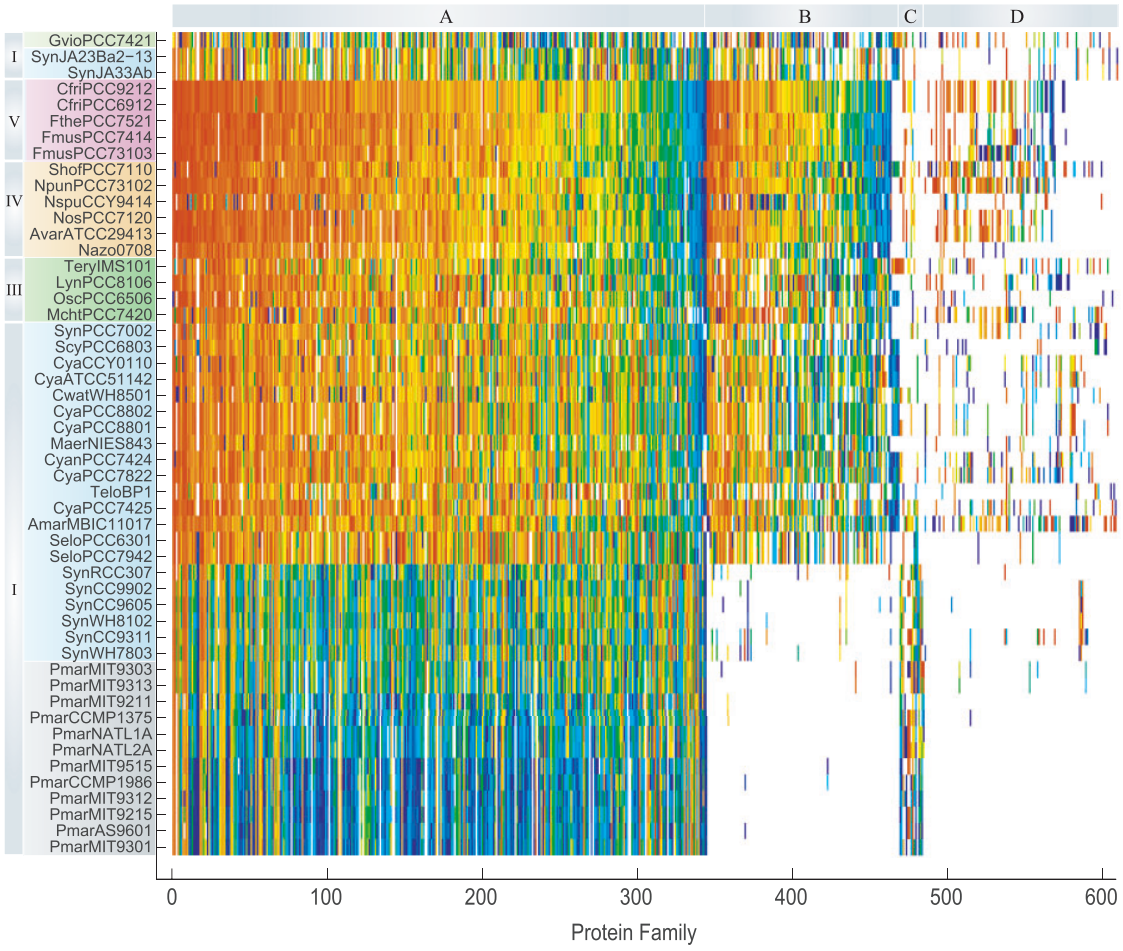


Fig. 4.—Presence/absence and sequence similarity patterns of cyanobacterial protein families by comparison with their homologs of endosymbiotic origin in six photosynthetic eukaryotes. Amino acid sequence similarity between the cyanobacterial proteins (*x* axis) and their counterparts in the eukaryotic plastid-derived set of protein families (*y* axis), as deduced for the genomes in the data set. Cell shades in the matrix correspond to the similarity ranking for each protein family (i.e., line) according to a color gradient from red (high similarity) to blue (low similarity). White cells correspond to genes lacking in the respective genomes. Protein families are ordered according to their distribution pattern into (*A*) nearly universal, (*B*) sparse representation or (*C*) highly frequent in the oceanic species, and (*D*) generally sparse representation. Cyanobacterial strains are ordered according to the MLN in fig. 2.

Therefore, we turned our attention to the larger set of nuclear genes of cyanobacterial origin whose homologs are not universally distributed among cyanobacteria. For 611 plant nuclear gene families identified as plastid acquisitions, we scored gene presence and absence, and protein sequence identity among cyanobacterial genomes (fig. 4). The SynPro clade lacks a substantial portion of these plastid ancestor gene families. A total of 245 (40%) protein families possessed by plants are absent in all *Prochlorococcus* strains, 137 (22%) are absent in all *Synechococcus* strains (fig. 4). The similarity map also shows that overall protein sequence similarity of plant nuclear genes is highest to homologs in members of subsection IV and V. For 225 (37%) protein families, the average amino acid identity between the cyanobacterial genes and their plant homologs is significantly higher for subsection V genomes ($\alpha = 0.05$, Kolmogorov–Smirnov test and FDR) than for subsection I genomes. When subsection IV and V genomes are combined and compared with those of subsection I, the value increases to 270 (44%) ($\alpha = 0.05$, Kolmogorov-Smirnov test and FDR). Thus, subsection IV and V genomes harbor more homologs of genes that plants acquired from cyanobacteria and those have higher sequence similarity to their plant homologs than genomes of subsection I. Similar amino acid usage in different organisms may sometimes lead to an overestimation of species relatedness (Rodríguez-Ezpeleta and Embley 2012). Here, we tested for such possible bias using a principle component analysis (PCA) for the amino acid frequencies encoded by the 611 genes of endosymbiotic origin. The transformation of amino acid usage into two principal components explains in total 89% of the variability observed

(supplementary fig. S7, Supplementary Material online). Furthermore, the PCA reveals that the eukaryotic species do not group with the filamentous cyanobacteria; hence, the protein sequence similarity observed between those two groups is not a result of biased amino acid usage. Consequently, we can conclude that in the present sample, the collection of genes possessed by the ancestor of plastids was most similar to that in filamentous, heterocyst-forming cyanobacteria (fig. 2).

Discussion

Possible Initial Benefits of Plastids

Today plastids supply fixed carbon to plant cells, but they also have a myriad of other functions in amino acid, lipid, and cofactor biosynthesis as well as nitrogen metabolism. What was the biochemical or physiological context of the symbiosis that gave rise to plastids—what initially associated the founder endosymbiont to its host in the first place? Traditional reasoning on the selective advantage that was crucial to the establishment of the plastid has it that the production of carbohydrates by the cyanobacterial endosymbiont was the key, a view that was clearly expressed by Mereschkowsky (1905, p. 605) in his initial formulation of endosymbiotic theory: "Plant cells receive with no effort whatsoever large amounts of preformed organic substrates (carbohydrates), which their chromatophores willingly supply."

An alternative suggestion is that the initial advantage of plastids may have simply been their uniquely useful metabolic end product, O_2, as a boost to respiration in early mitochondria (Martin and Müller 1998). The chemical benefit of O_2 could, of course, have only been of value if the initial endosymbiosis had taken place at a time in Earth's history, or in an environment, where O_2 was not freely available in sufficient amounts. Fossil evidence supports the notion that the primary plastid endosymbiosis occurred at least 1.2 billion years ago (Butterfield 2000) and molecular estimates suggest that plastids might have arisen by approximately 1.5 billion years ago (Parfrey et al. 2011). Geochemists have found over the last decade that an approximately 2 billion year span of protracted ocean anoxia ended only about 580 Ma (Anbar and Knoll 2002; Johnston et al. 2009; Lyons et al. 2009; Lyons and Reinhardt 2009; Sahoo et al. 2012). The six major eukaryotic assemblages or "supergroups" currently recognized, including plants, arose and diversified during that time (Parfrey et al. 2011), that is, while the oceans were still anoxic (Müller et al. 2012). Such geological context (ocean anoxia during most of the Proterozoic) would be compatible with a possible role for O_2 as an initial benefit in the plastid evolution. Indeed, for Stanier (1970), the production of O_2 was a reason to suggest that plastids arose before mitochondria did. Of course, Proterozoic ocean anoxia was likely less pronounced in the photic zone than below it (Johnston et al. 2009). A freshwater origin of plastids is also a possibility to consider, whereby the present data linking plastids phylogenomically more closely with freshwater cyanobacteria than with marine forms (fig. 2) would be compatible with that view.

Another suggestion is that the key to establishment of the plastid was the origin of carbon translocators in the plastid inner membrane and that the incorporation of a metabolite antiporter like the triose phosphate translocator in the ancestral plastid membrane was the essential step for establishing the primary endosymbiosis by allowing the plant ancestor to profit from cyanobacterial carbon fixation (Weber et al. 2006). In the same vein, it was furthermore argued that the key to establishment of the plastid entailed the insertion of additional host-controlled metabolite exchange proteins into plastid membranes fulfilling a similar export role (Gross and Bhattacharya 2009). A problem with theories that focus on carbon exporters as the key innovation at plastid origin is that cyanobacteria are well known to produce copious amounts of exopolysaccharides (De Philippis and Vincenzini 1998), such that there would be no need to evolve or insert transporters for provision of carbohydrates to be realized by the host.

The theory for the initial benefit of plastids that is currently best founded in direct observation, we would argue, is that nitrogen fixation was a key to the establishment of the symbiosis (Kneip et al. 2007). This view is supported by the circumstance that in modern symbioses involving cyanobacteria, nitrogen (not reduced carbon) is usually the key nutrient underlying the success of the partnership (Rai et al. 2000; Raven 2002). Accordingly, the cyanobacterial endosymbionts are nitrogen fixing forms and combined nitrogen (ammonium) is the nutrient provided by the cyanobacterium. This is true for diatoms with N_2-fixing cyanobacterial endosymbionts (Prechtl et al. 2004; Kneip et al. 2008), prymnesiophytes with associated N-fixing cyanobacteria that might be ectosymbionts (Thompson et al. 2012), cyanobionts in lichens (Rikkinen et al. 2002), coralloid roots of cycads (Costa et al. 2004), the angiosperm *Gunnera* (Chiu et al. 2005), and the water-fern *Azolla* (Ran et al. 2010). In the case of *Azolla* and *Rhopalodia*, the N_2-fixing cyanobacteria live as intracellular endosymbionts (Kneip et al. 2008; Ran et al. 2010).

Recent studies have suggested that a filamentous phenotype and heterocyst differentiation may have been hallmark phenotypic characteristics of the plastid ancestor (Deusch et al. 2008; Ran et al. 2010; Larsson et al. 2011). Indeed, in modern cyanobacterial symbioses, fixed nitrogen is the main currency of benefit that the cyanobacterial symbiont provides to its host (Kayley et al. 2007). The early physiological association of the plastid ancestors with their host might thus have been similar to that of the unicellular nitrogen-fixing endosymbiont and its diatom host *Rhopalodia* (Kneip et al. 2008), or the highly reduced *Nostoc azollae*, an obligate cyanobiont of water-ferns, whose genome has drastically been reduced,

with a large portion of the remaining genes specifically dedicated to heterocyst differentiation and nitrogen fixation (Ran et al. 2010). A potential problem with this view is that nitrogen fixation has not been retained by any modern plant (Allen and Raven 1996). Why not? One possible reason concerns the circumstance that cyanobacterial O_2 production led to an oxidation state of the environment in which nitrate became abundant (Falkowski et al. 2008)—in a world of abundant nitrate, nitrogenase is less necessary, hence less likely to be retained, although one should recall that modern cyanobacterial symbionts do fix nitrogen for their hosts. Perhaps, more importantly, in oxic environments cyanobacteria that express nitrogenase must exhibit either temporal separation of photosynthesis and nitrogen fixation (N_2-fixation occurring mainly in the dark; Mitsui et al. 1986), or other means of protecting the notoriously O_2-sensitive enzyme from inactivation such as diazocyte differentiation in *Trichodesmium* (Sandh et al. 2012), or heterocyst formation in subsections IV and V (Kumar et al. 2010). It is possible that such nitrogenase-protecting strategies, whereas readily accessible to genetically autonomous prokaryotes, are not among the realm of possibilities that plastids, which relinquished most of their genetic autonomy, can developmentally attain.

Many Endosymbionts, or Only One with Many Genes?

Gene transfer following plastid origin readily explains plant nuclear genes that branch with cyanobacteria. However, many plant-specific genes branch with other prokaryotes (fig. 5*A*). Plant genes that branch with chlamydial homologs have led to the inferences that a chlamydial endosymbiont accompanied the origin of plastids (Brinkman et al. 2002; Huang and Gogarten 2007; Price et al. 2012). This theory postulates that the plant ancestor consumed cyanobacteria as food and was parasitized by environmental chlamydias (Huang and Gogarten 2007; Moustafa et al. 2008), whereby the chlamydias were key to establishing the plastid because chlamydia-like bacteria donated genes that allowed export of photosynthate from the cyanobacterial plastid ancestor and its polymerization into storage polysaccharide in the cytosol (Price et al. 2012). The flaw with this theory is that it is based on the uncritical interpretation of computational results of genome comparisons that, as has long been known (Rujan and Martin 2001; Martin et al. 2002; Esser et al. 2007; Dagan et al. 2008), would implicate many other groups of prokaryotes far more strongly than they would implicate chlamydias as active bystanders at the origin of plastids. The focus on chlamydia as opposed to, say spirochaetes or proteobacteria, is arbitrary and to some extent ad hoc. If one were to take the chlamydia theory seriously, or think it through in full, the transiently symbiotic and gene-dealing "chlamydioplast" would have to take a number and wait in line next to the actinobacterioplast, the clostridioplast, the bacilloplast, the bacteriodetoplast, and the spirochaetoplast, and so forth. (fig. 5*A*). Beyond the

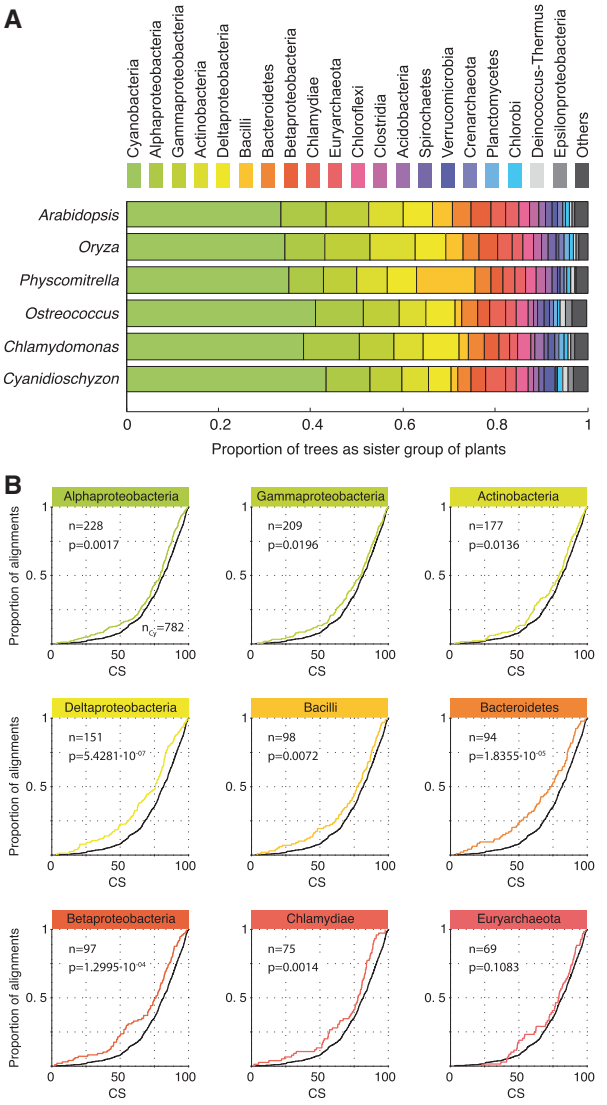


Fig. 5.—Taxon distribution of nearest neighbors to plant genes. (*A*) Tree samples distribute as following: *Arabidopsis*: 2,324; *Oryza*: 1,792; *Physcomitrella*: 2,511; *Ostreococcus*: 968; *Chlamydomonas*: 1,218; and *Cyanidioschyzon*: 693. Microbial taxonomic groups having a low frequency of nearest neighbors were grouped into the "Others" bar. Those include Aquificae, Dictyoglomi, Elusimicrobia, Fibrobacteres, Fusobacteria, Gemmatimonadetes, Korarchaeota, Nanoarchaeota, Nitrospirae, Tenericutes, Thaumarchaeota, and Thermotogae. (*B*) A comparison of alignment quality (CS) between trees of *Arabidopsis* genes having a cyanobacterial nearest neighbor (black) and trees where a nearest neighbor from a different prokaryotic group was inferred (colored according to the taxa). In all groups but the Euryarchaeota, the alignment quality of trees where a noncyanobacterial nearest neighbor was inferred is significantly lower in comparison with tree topologies having cyanobacteria as their nearest neighbor (using Wilcoxon test, $\alpha = 0.05$). These results suggest that the inference of noncyanobacterial nearest neighbors to plant genes is less reliable than the inference of cyanobacterial nearest neighbors.

cyanobacterial signal, which corresponds to a tangible double membrane-bounded and DNA-containing organelle, the other putative phylogenetic signals in the data, especially that involving chlamydia, are better explained in terms of known phenomena, such as LGT among free-living prokaryotes (Dagan et al. 2008) and by phylogeny reconstruction errors (White et al. 2007; Stiller 2011) (fig. 5*B*), both of which we know to really exist, than in terms of gene dealing endosymbionts whose existence is inferred from a few gene trees. The null hypothesis for endosymbiotic theory in the age of genomes should be: The ancestors of plastids underwent LGT, just like modern cyanobacteria, whose genomes are chimeras of genes from many sources (Mulkidjanian et al. 2006), and the plastid ancestor genome was probably no different (Richards and Archibald 2011). LGT among prokaryotes accounts for the diverse sequence affinities of genes acquired from the single ancestor of plastids with far fewer corollaries than a one-symbiont-per-gene theory. We merely need to incorporate the effect that LGT among prokaryotes will have over geological time on the endosymbiotic origins of organelles.

Clues to the Origin of Two Photosystems

One notable aspect of cyanobacterial phylogenomics presented in this study is that the marine cyanobacteria are not basal in the trees (fig. 2 and supplementary fig. S3, Supplementary Material online). These small unicellular cyanobacteria (diameter 1 μm or less) share reduced genome sizes (<3 Mb) as a common trait, and seem to have arisen from ancestors with larger genomes (Larsson et al. 2011) that, inferred from the phylogeny, lived in terrestrial, brackish, or perhaps freshwater environments (Sánchez-Baracaldo et al. 2005). This led Blank and Sánchez-Baracaldo (2010) to suggest that oxygenic photosynthesis arose in a freshwater environment. Our results support that view, and this conclusion has implications for the origin of water-splitting photosynthesis. Among many possibilities (Xiong and Bauer 2002; Hohmann-Marriot and Blankenship 2011; Williamson et al. 2011), it has been suggested that the progenitor of the cyanobacteria had genes for both type I (RCI) and type II (RCII) photosynthetic reaction centers (via gene duplication) but expressed either set of genes depending on the reducing conditions in the environment (Allen 2005): type RCI in the presence of H_2S for noncyclic electron flow, as in *Chlorobium* (or the facultative anaerobic cyanobacterium *Oscillatoria limnetica*); and type RCII in the absence of H_2S, for cyclic electron flow, as in *Rhodobacter* (Allen 2005). Were regulation to fail such that both type I and type II reaction centers became expressed in the absence of H_2S, the protocyanobacterium would oxidatively perish, unless it could extract electrons from an environmentally available donor.

Such an electron donor could have been aqueous $Mn^{II/III}$, which has the utilitous property of being photo-oxidized by ultraviolet light (Allen and Martin 2007), an abundant component of solar radiation incident on the Earth's surface prior to accumulation of atmospheric oxygen. Attaining suitably high concentrations of $Mn^{II/III}$ as an environmentally available electron donor in the ocean would be problematic, but not in a freshwater setting. Allen et al. (2012) have recently shown that an engineered, Mn-binding type II reaction center of *Rhodobacter sphaeroides* will produce O_2 from O_2^- in the presence of Mn in a light-dependent reaction in which photo-damage is impeded in comparison with that in a wild-type, Mn-free reaction center. Their observation (Allen et al. 2012) is likely an important clue to the origin of oxygenic photosynthesis, at which time a protocyanobacterial type II reaction center acquired, via natural selection, the ability to (photo-)oxidize $Mn^{II/III}$—itself ultimately rereduced by water—and then to reduce a newly constitutive type I reaction center. Transition from environmental (substrate) $Mn^{II/III}$ ions to the catalytic Mn_4Ca center of cyanobacterial RCII would then have permitted light-dependent CO_2 and/or nitrogen fixation, in the absence of electron donors other than water.

What Makes a Branching Cyanobacterium?

The morphological diversity of cyanobacteria poses an intriguing question in the biology and evolution of cell differentiation. Transposon mutagenesis of *Synechococcus elongatus* PCC 7942 (subsection I) revealed that the loss of several genes involved in cell division leads to filament formation (Miyagishima et al. 2005). However, our analysis revealed that all recognized cyanobacterial cell division genes are present in the genomes of filamentous cyanobacteria, including those of subsection V. This suggests that the filamentous phenotype in cyanobacteria of subsections III, IV, and V is not due to loss of genes for cell division, though it is currently unknown whether those that are present are all expressed. Genes common to both unicellular and filamentous cyanobacteria may also be important for determining trichome structure in members of subsections III–V. This is suggested by a recent study on the filamentous heterocystous strain *N. punctiforme* ATCC 29133 (Lehner et al. 2011), which showed that mutations of the *amiC2* gene, encoding an amidase involved in septa formation, will lead to a morphology similar to that of colonial unicellular cyanobacteria, and prevent heterocyst differentiation. Furthermore, filament formation in *S. elongatus* PCC 7942 can be induced by over-expression of the gene encoding FtsZ, which is known as a cell division protein (Mori and Johnson 2001). Thus, the lack of clear candidate genes whose distribution across cyanobacterial genomes correlate with cellular morphology and the experimental evidence that links between the expression level (rather than presence/absence) of cell division proteins and filament formation suggest that a filamentous

phenotype may result from modifications of the gene regulatory network and cell division program.

Supplementary Material

Supplementary figures S1–S7 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The work in the authors' laboratories is supported by SFB-TR1 to T.D. and W.F.M., the European Research Council (grant no. 232975 to W.F.M.; grant no. 281357 to T.D.), and a Leverhulme Trust Research Grant (no. F07 476AQ to J.F.A.). The support by the Institut Pasteur and the Centre National de la Recherche Scientifique (URA 2172) is acknowledged by M.G., R.R., and N.T.M. The authors are grateful to T. Coursin and T. Laurent for technical assistance in maintaining the Pasteur Culture Collection of Cyanobacteria at the Institut Pasteur. Additional computational support and infrastructure was provided by "Zentrum fuer Informations- und Medientechnologie" (ZIM) at Heinrich-Heine-University, Duesseldorf, Germany.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.
Allen JF. 2005. A redox switch hypothesis for the origin of two light reactions in photosynthesis. FEBS Lett. 579:963–938.
Allen JF, Martin W. 2007. Evolutionary biology: out of thin air. Nature 445: 610–612.
Allen JF, Raven JA. 1996. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. J Mol Evol. 42:482–492.
Allen JP, et al. 2012. Light-driven oxygen production from superoxide by Mn-binding bacterial reaction centers. Proc Natl Acad Sci U S A. 109: 2314–2318.
Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 35: 3389–3342.
Anbar AD, Knoll AH. 2002. Proterozoic ocean chemistry and evolution: a bioinorganic bridge. Science 297:1137–1142.
Badger JH, Olsen GJ. 1999. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol. 16:512–524.
Baymann F, Brugna M, Mühlenhoff U, Nitschke W. 2001. Daddy, where did (PS)I come from? Biochim Biophys Acta. 1507:291–310.
Bekker A, et al. 2004. Dating the rise of atmospheric oxygen. Nature 427: 117–120.
Blank CE, Sánchez-Baracaldo P. 2010. Timing of morphological and ecological innovations in the cyanobacteria—a key to understanding the rise in atmospheric oxygen. Geobiology 8:1–23.
Bowes JB, et al. 2008. Xenbase: a *Xenopus* biology and genomics resource. Nucleic Acids Res. 36:D761–D767.
Brinkman FS, et al. 2002. Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. Genome Res. 12:1159–1167.
Butterfield NJ. 2000. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the mesoproterozoic/neoproterozoic radiation of eukaryotes. Paleobiology 263:386–404.
Chiu WL, et al. 2005. Nitrogen deprivation stimulates symbiotic gland development in *Gunnera manicata*. Plant Physiol. 139:224–230.
Costa JL, Romero EM, Lindblad P. 2004. Sequence based data supports a single *Nostoc* strain in individual coralloid roots of cycads. FEMS Microbiol Ecol. 49:481–487.
Criscuolo A, Gribaldo S. 2011. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. Mol Biol Evol. 28:3019–3032.
Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105:10039–10044.
De Philippis R, Vincenzini M. 1998. Exocellular polysaccharides from cyanobacteria and their possible applications. FEMS Microbiol Rev. 22:151–175.
Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27: 4636–4641.
Deschamps P, et al. 2008. Metabolic symbiosis and the birth of the plant kingdom. Mol Biol Evol. 25:536–548.
Deusch O, et al. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 25:748–761.
Doolittle WF. 1980. Revolutionary concepts in evolutionary biology. Trends Biochem Sci. 5:146–149.
Doolittle WF, Bapteste E. 2007. Pattern pluralism and the tree of life hypothesis. Proc Natl Acad Sci U S A. 104:2043–2049.
Dufresne A, et al. 2008. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol. 9:R90.
Enright AJ, van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30: 1575–1584.
Esser C, Martin W, Dagan T. 2007. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 3:180–184.
Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034–1039.
Felsenstein J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Seattle (WA): University of Washington.
Franche C, Damerval T. 1988. Test on nif probes and DNA hybridizations. Methods Enzymol. 167:803–808.
Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195–202.
Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. Annu Rev Plant Biol. 59:491–517.
Gross J, Bhattacharya D. 2009. Opinion: Mitochondrial and plastid evolution in eukaryotes: an outsiders' perspective. Nat Rev Genet. 10:495–505.
Gross J, Meurer J, Bhattacharya D. 2008. Evidence of a chimeric genome in the cyanobacterial ancestor of plastids. BMC Evol Biol. 8:117.
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.
Hohmann-Marriott MF, Blankenship RE. 2011. Evolution of photosynthesis. Annu Rev Plant Biol. 62:515–548.
Huang J, Gogarten JP. 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. 8:R99.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.
Johnston DT, Wolfe-Simon F, Pearson A, Knoll AH. 2009. Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. Proc Natl Acad Sci U S A. 106:16925–16929.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation rate matrices from protein sequences. Comput Appl Biosci. 8:275–282.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Kayley MU, Bergman B, Raven JA. 2007. Exploring cyanobacterial mutualism. Annu Rev Ecol Evol Syst. 38:255–273.

Kneip C, Lockhart P, Voss C, Maier UG. 2007. Nitrogen fixation in eukaryotes—new models for symbiosis. BMC Evol Biol. 7:55.
Kneip C, Voss C, Lockhart PJ, Maier UG. 2008. The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. BMC Evol Biol. 8:30.
Krause L, et al. 2007. GISMO-gene identification using a support vector machine for ORF identification. Nucleic Acids Res. 35:540–549.
Kumar K, Mella-Herrera RA, Golden JW. 2010. Cyanobacterial heterocysts. Cold Spring Harb Perspect Biol. 2:a000315.
Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.
Larsson J, Nylander JA, Bergman B. 2011. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evol Biol. 11:187.
Lehner J, et al. 2011. The morphogene AmiC2 is pivotal for multicellular development in the cyanobacterium *Nostoc punctiforme*. Mol Microbiol. 79:1655–1669.
Lindell D, et al. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A. 101: 11013–11018.
Lockhart P, et al. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol. 23:40–45.
Lyons TW, Anbar AD, Severmann S, Scott C, Gill BC. 2009. Tracking euxinia in the ancient ocean: a multiproxy perspective and Proterozoic case study. Annu Rev Earth Planet Sci. 37:507–534.
Lyons TW, Reinhard CT. 2009. An early productive ocean unfit for aerobics. Proc Natl Acad Sci U S A. 106:18045–18046.
Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37–41.
Martin W, et al. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A. 99: 12246–12251.
Matsuzaki M, et al. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428:653–657.
Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Centralbl. 25:593–604. [English translation in Eur J Phycol. 1999;34:287–295.]
Meyer F, et al. 2003. GenDB—an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31:2187–2195.
Mitsui AS, et al. 1986. Strategy by which nitrogen-fixing unicellular cyanobacteria grow photoautotrophically. Nature 323:720–722.
Miyagishima SY, Wolk CP, Osteryoung KW. 2005. Identification of cyanobacterial cell division genes by comparative and mutational analyses. Mol Microbiol. 56:126–143.
Moisander PH, et al. 2010. Unicellular cyanobacterial distributions broaden the oceanic N2 fixation domain. Science 327:1512–1524.
Morel FM, Price NM. 2003. The biogeochemical cycles of trace metals in the oceans. Science 300:944–947.
Mori T, Johnson CH. 2001. Independence of circadian timing from cell division in cyanobacteria. J Bacteriol 183:2439–2444.
Moustafa A, Reyes-Prieto A, Bhattacharya D. 2008. Chlamydiae has contributed at least 55 genes to plantae with predominantly plastid functions. PLoS One 3:e2205.
Mulkidjanian AY, et al. 2006. The cyanobacterial genome core and the origin of photosynthesis. Proc Natl Acad Sci U S A. 103: 13126–13131.
Müller M, et al. 2012. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol Mol Biol Rev. 76:444–495.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.
Palenik B, et al. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A. 104:7705–7710.
Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci U S A. 108:13624–13629.
Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier UG. 2004. Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. Mol Biol Evol. 21:1477–1481.
Price DC, et al. 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. Science 335:843–847.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.
Rai AN, Söderbäck E, Bergman B. 2000. Cyanobacterium-plant symbioses. New Phytol. 147:449–481.
Ran L, et al. 2010. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS One 5:e11486.
Raven JA. 2002. Evolution of cyanobacterial symbioses. In: Rai AN, Bergman B, Rasmussen U, editors. Cyanobacteria in symbiosis. Dordrecht (The Netherlands): Kluwer Academic Publishers. p. 326–246.
Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE. 2002. Whole-genome analysis of photosynthetic prokaryotes. Science 298:1616–1620.
Reyes-Prieto A, et al. 2010. Differential gene retention in plastids of common recent origin. Mol Biol Evol. 27:1530–1537.
Richards TA, Archibald JM. 2011. Cell evolution: gene transfer agents and the origin of mitochondria. Curr Biol. 21:R112.
Richly E, Leister D. 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. Gene 329:11–16.
Rikkinen J, Oksanen I, Lohtander K. 2002. Lichen guilds share related cyanobacterial endosymbionts. Science 297:357.
Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. J Gen Microbiol. 111:1–61.
Rippka R, Herdman H. 2002. Pasteur culture collection of Cyanobacteria: catalogue and taxonomic handbook. I. Catalogue of strains. Paris: Institut Pasteur.
Rocap G, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042–1047.
Rodríguez-Ezpeleta N, Embley TM. 2012. The SAR11 group of alphaproteobacteria is not related to the origin of mitochondria. PLoS One 7:e30520.
Rujan T, Martin W. 2001. How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies. Trends Genet. 17:113–120.
Sahoo SK, et al. 2012. Ocean oxygenation in the wake of the Marinoan glaciation. Nature 489:546–549.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Sánchez-Baracaldo P, Hayes PK, Blank CE. 2005. Morphological and habitat evolution in the Cyanobacteria using a compartmentalization approach. Geobiology 3:145–165.
Sandh G, Xu Linghua, Bergman B. 2012. Diazocyte development in the marine diazotrophic cyanobacterium *Trichodesmium*. Microbiology 158:345–352.
Sharma AD, Gill PK, Singh P. 2002. DNA isolation from dry and fresh samples of polysaccharide-rich plants. Plant Mol Biol Rep. 20: 415a–415f.
Sharon I, et al. 2009. Photosystem I gene cassettes are present in marine virus genomes. Nature 461:258–262.
Shi T, Falkowski PG. 2008. Genome evolution in cyanobacteria: the stable core and the variable shell. Proc Natl Acad Sci U S A. 105:2510–2515.
Stanier RY. 1970. Some aspects of the biology of cells and their possible evolutionary significance. Symp Soc Gen Microbiol. 20:1–38.

Genome Biology and Evolution SMBE

Stiller JW. 2011. Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. BMC Evol Biol. 11:259.
Stiller JW, Hall BD. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol Biol Evol. 16:1270–1279.
Stucken K, et al. 2010. The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. PLoS One 5:e9235.
Swingley WD, et al. 2008. Niche adaptation and genome expansion in the chlorophyll *d*-producing cyanobacterium *Acaryochloris marina*. Proc Natl Acad Sci U S A. 105:2005–2010.
Tatusov RL, et al. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29:22–28.
Thompson AW, et al. 2012. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. Science 337:1546–1550.
van Mooy BA, et al. 2009. Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. Nature 458: 69–72.
Weber AP, Linka M, Bhattacharya D. 2006. Single, ancient origin of a plastid metabolite translocator family in Plantae from an endomembrane-derived ancestor. Eukaryot Cell. 5:609–612.
Whelan S, Goldman N. 2001. A general model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.
White WT, Hills SF, Gaddam R, Holland BR, Penny D. 2007. Treeness triangles: visualizing the loss of phylogenetic signal. Mol Biol Evol. 24:2029–2039.
Williamson A, Conlan B, Hillier W, Wydrzynski T. 2011. The evolution of photosystem II: insights into the past and future. Photosynth Res. 107: 71–86.
Xiong J, Bauer CE. 2002. Complex evolution of photosynthesis. Annu Rev Plant Biol. 53:503–521.
Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. 2009. Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. Genome Biol Evol. 1:325–339.
Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16:1099–1108.

Associate editor: John Archibald

6 Zusammenfassung der Ergebnisse

6.1 Genome networks root the tree of life between prokaryotic domains

Die Wurzel eines phylogenetischen Baumes gibt die Richtung und die Abfolge der evolutionären Ereignisse zwischen den untersuchten Taxa und ihren hypothetischen Vorfahren vor. Die genaue Kenntnis über die Position der Wurzel im phylogenetischen Stammbaum aller Lebewesen auf der Erde ist wichtig, um fundamentale Rückschlüsse auf die Entstehung des Lebens ziehen zu können (Lake, 2009). Die Position dieser Wurzel wird heute jedoch immer noch kontrovers diskutiert. Einige Analysen leiteten die Wurzel und somit die älteste Aufteilung aller Lebewesen innerhalb der Eubakterien ab (Cavalier-Smith, 2006, 2009, 2010a,b; de Duve, 2007). Dieses Szenario impliziert jedoch gleichzeitig eine erhöhte archaebakterielle Evolutionsrate, um die substantiellen molekularen Unterschiede zwischen Archaebakterien und Eubakterien zu erklären (Cavalier-Smith, 2010b). Dieser Teil der Arbeit beschäftigte sich mit der Frage, ob es tatsächlich Anzeichen für eine solche relative Erhöhung der Evolutionsrate unter den Archaebakterien gibt. Die üblichen Methoden, eine Wurzel im phylogenetischen Baum zu bestimmen, haben mit Problemen der tiefen Genphylogenie (Philippe *et al.*, 2011), sowie des teilweise massiven lateralen Gentransfers unter Prokaryoten zu kämpfen (Doolittle, 1999). Dies führte zu weitgehenden Konflikten in den Ergebnissen. Bei der Bestimmung der tiefsten und damit ältesten Aufteilung im Stammbaum aller Lebewesen wurde hier deshalb ein alternatives Verfahren angewandt, welches vollständige Genomdaten verwendete und unabhängig von der Rekonstruktion phylogenetischer beziehungsweise phylogenomischer Bäume war. Gleichzeitig konnten auch solche Gene, die in ihrer Geschichte lateralem Transfer unterlagen, in die Analyse mit einbezogen werden, ohne diese gemäß üblicher Praxis zuvor aussortieren zu müssen.

Mit Hilfe von in 191 prokaryotischen Genomen kodierten Proteinsequenzdaten konnte gezeigt werden, dass sich mit Hilfe unterschiedlicher Schwellenwerte für die paarweise Aminosäuresequenzidentität Proteinfamilien erzeugen lassen, aus deren Vergleich unter benachbarten Schwellenwerten Genomaufteilungsinformationen (engl. *splits*) unterschiedlicher Altersklassen ermittelt werden konnten. Dies ermöglichte es, phylogenetische Netzwerke für die unterschiedlichen Altersklassen direkt aus den so gewonnenen Genomaufteilungsinformationen zu erstellen. Für das phylogenetische Netzwerk der ältesten Aufteilungs-Kategorie wurde das *Midpoint-rooting*-Verfahren (Farris, 1972) auf die Verwendung in phylogenetischen Netzwerken angepasst und angewendet. Die ermittelte Wurzel teilt die beiden prokaryotischen Domänen der Archaebakterien und der Eubakterien. Mit Hilfe eines Jacknife-Verfahrens konnte diese Position als robust eingestuft werden.

Eine Analyse von Quartets paraloger Sequenzen, welche im letzten gemeinsamen Vorfahren der Prokaryoten dupliziert wurden (Kollman und Doolittle, 2000) zeigte, dass die betrachteten Gene einer nur geringen lateralen Transferrate zwischen den höheren taxonomischen Gruppen unterlagen. Für 75 Prozent dieser Quartets wurde eine globale Evolutionsrate bestimmt, während für 19 Prozent linienspezifische Evolutionsraten festgestellt werden konnten. Letztere erlaubten eine Einteilung in langsam bis schnell evolvierende taxonomischen Gruppen. Für die Gruppe der archaebakteriellen Sequenzen ergab sich hierbei nur eine mittlere Evolutionsrate. Das Szenario einer erhöhten Evolutionsrate der Archaebakterien konnte damit ausgeschlossen werden. Lediglich die Linien der γ-Proteobakterien, der α-Proteobakterien, der Aktinobakterien sowie der Bacilli wiesen erhöhte relative Evolutionsraten auf.

Diese Arbeit unterstreicht, die Antiquität der Archaebakterien, welche mindestens genau so alt sind wie die Eubakterien.

6.2 A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies

Dieser Teil der Arbeit beschäftigte sich mit der Ableitung von Ereignissen lateralen Gentransfers mit Hilfe der phylogenetischen Methode. Bei diesem Verfahren werden phylogenetische Bäume einzelner Gene beziehungsweise Proteine rekonstruiert und mit einer Referenz, dem Stammbaum aller betrachteten Spezies,

verglichen. Treten zwischen diesen beiden Topologien inkompatible Äste auf und sind diese zusätzlich statistisch hinreichend stark unterstützt, kann davon ausgegangen werden, dass in der Evolutionsgeschichte des entsprechenden Gens ein lateraler Transfer stattgefunden hat. Es wurde untersucht, ob es Anzeichen für statistisch hinreichend unterstützte Phylogenien gibt, welche dennoch inkorrekt sind. Diese spiegeln dann nicht die tatsächlich stattgefundenen Evolutionsereignisse wider und führen so zu falschen Aussagen bezüglich Art und Ausmaßes lateralen Gentransfers. Im speziellen sollte hier der Einfluss des Sequenzalignments auf die Ableitung divergenter Phylogenien analysiert werden.

Anhand der Analyse von Ergebnissen einer bereits veröffentlichten Studie, welche sich mit der Ableitung von lateralen Gentransferereignissen beschäftigte, konnte ein starker Einfluss des Sequenzalignments auf die getroffenen Schlussfolgerungen festgestellt werden. Alignments, welche in der phylogenetischen Rekonstruktion hinreichend stark unterstützte inkompatible Äste erzeugt hatten, wiesen jeweils Eigenschaften auf, welche die Erzeugung inkorrekter phylogenetischer Bäume begünstigen. Dies zeigte sich in einer erhöhten Anzahl an beinhalteten Sequenzen, einer größeren Sequenzdivergenz, einer geringeren Anzahl informativer Postitionen sowie der Erhöhung von Artefakten aufgrund des Alignmentverfahrens. Mit Hilfe einer Support-Vektor-Maschine konnte allein aus den Eigenschaften der Alignments mit nahezu 80-prozentiger Sicherheit bestimmt werden, ob diese während der phylogenetischen Rekonstruktion inkompatible Äste mit ausreichend hoher statistischer Unterstützung erzeugen werden. Hierbei waren die für die Klassifizierung entscheidenden Parameter die Sequenzvariabilität, Alignmentverlässlichkeit, sowie die Unterscheidung, ob die untersuchten Alignments Sequenzen aus unterschiedlichen Reichen (Eubakterien, Archaebakterien oder beide zusammen) enthielten.

Man kann davon ausgehen, dass die Verschiebung der Alignmenteigenschaften in Richtung einer größeren Wahrscheinlichkeit zur Manifestierung phylogenetischer Artefakte, eine verlässliche Rekonstruktion phylogenetischer Bäume verhindert. Inkorrekt rekonstruierte Äste dürften also eine nicht zu vernachlässigende Quelle für die Ableitung von Ereignissen lateralen Gentransfers darstellen. Die Ergebnisse deuten darauf hin, dass bei problematischen Alignments von geringerer Qualität sehr viel häufiger unstimmige Äste allein aufgrund von phylogenetischen Rekonstruktionsartefakten auftreten, als dies für weniger problematische Alignments der Fall ist.

6.3 Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids

Auf der Basis früherer Untersuchen mit widersprüchlichen Ergebnissen (Criscuolo und Gribaldo, 2011; Deschamps *et al.*, 2008; Deusch *et al.*, 2008; Reyes-Prieto *et al.*, 2010) und aktueller Genomdaten beschäftigte sich dieser Teil der Arbeit mit der Frage, welche frei lebenden rezenten Cyanobakterien einem hypothetischen Vorfahren des Plastiden am ähnlichsten sind. Es sollte abgeschätzt werden, wie groß der Anteil in den Kerngenomen heutiger photosynthetischer Algen und Pflanzen ist, der sich auf die Folgen einer primären Endosymbiose zurück führen lässt. Hierbei sollte vor allem auch der Einfluss der Qualität des Sequenzalignments auf die erzielten Ergebnisse betrachtet werden. Zusätzlich wurden vertikal und lateral vererbte Komponenten der cyanobakteriellen Evolution untersucht. Dabei lag der besondere Fokus auf einem möglichen Ursprung der oxygenen Photosynthese in einer Süßwasserumgebung. In die Analyse flossen Genomsequenzierungsdaten von fünf Vertretern (*Chlorogloeopsis fritschii* PCC 6912, *C. fritschii* PCC 9212, *Fischerella muscicola* PCC 7414, *F. muscicola* PCC 73103, *F. thermalis* PCC 7521) einer in Genomdaten bisher nicht vertretenen Gruppe (Untergruppe V) der Cyanobakterien ein. Diese stellen morphologisch aufgrund ihres filamentösen Wachstums mit echten Verzweigungen und der Differenzierung spezialisierter Zellen zur Stickstofffixierung (Heterozysten) und Überdauerung (Akineten) sowie durch die Bildung eines beweglichen Filaments (Hormogonien), die komplexesten Cyanobakterien und damit auch eine der komplexesten Gruppe der Prokaryoten überhaupt dar. Zusätzlich wurde das Genom des filamentös mit Pseudoverzweigungen wachsenden, Heterozysten-bildenden Cyanobakteriums *Scytonema hofmanni* PCC 7110 (Untergruppe IV) sequenziert und analysiert. *S. hofmanni* ist mit 12.356 Protein-kodierenden Genen zur Zeit der Gen-reichste Prokaryot.

Es konnte gezeigt werden, dass dem Phänotyp der echten Verzweigungen, welcher die Untergruppe V im Gegensatz zu allen anderen Cyanobakterien auszeichnet, keine größere Stammmenge an Proteinen zugrunde liegt. Naheliegend war deshalb, dass nur wenige exprimierte Proteine, welche hauptsächlich die Regulation von Genen der Zellteilung und der Lokalisierung ihrer Produkte betreffen, von Bedeutung sind. Eine cyanobakterielle Phylogenie von 324 Proteinfamilien, deren Mitglieder in Einzelkopie pro Bakterienstamm vorlagen, legte einen gemein-

samen Ursprung aller filamentösen Cyanobakterien nahe und ließ die Entstehung der oxygenen Photosynthese in einer Süßwasserumgebung vermuten. Vergleiche dieses cyanobakteriellen Stammbaums mit einem phylogenetischen Netzwerk ergaben, dass diese Phylogenie nur sehr gering von Ereignissen lateralen Gentransfers betroffen war. Es konnte somit als Referenz zur Analyse der lateralen Komponente mit Hilfe eines minimalen lateralen Netzwerks (Dagan, Artzy-Randrup und Martin, 2008) dienen. Für zwei Drittel der Proteinfamilien wurde mindestens ein laterales Transferereignis abgeleitet, mit der Tendenz einer höheren Transferrate für Gene großer cyanobakterieller Genome. Die Übertragung von zehn oder mehr Genen stellte sich als äußerst seltenes Ereignis heraus. Zusätzlich wurden Anzeichen für eine Barriere des lateralen Gentransfers zwischen der Untergruppe I und den restlichen Cyanobakterien festgestellt.

Basierend auf 35.862 phylogenetischen Bäumen kernkodierter Algen- und Pflanzenproteine sowie ihrer Homologe in Prokaryoten und nicht-photosynthetischen Eukaryoten wurde untersucht, wie häufig Pflanzen und Cyanobakterien durch einen gemeinsamen Ast von den restlichen Organismen abgeteilt werden. Dabei zeigte sich ein großer Anteil von Genen endosymbiotischen Ursprungs in den untersuchten Algen und Pflanzen mit vorwiegenden Funktionen im Energie- und Kohlenstoffwechsel. Die unizellulären Algen wiesen die größten Anteile von Genen plastidären Ursprungs auf. Es zeigte sich, dass der abgeleitete Anteil an entdeckten transferierten Genen für unverlässliche Alignments geringer wird. Dies unterstrich ein weiteres Mal die zentrale Bedeutung der Qualität multipler Sequenzalignments in der molekularen Phylogenomik.

Der Versuch, mit Hilfe universeller kernkodierter Proteine cyanobakteriellen Ursprungs in Algen und Pflanzen mit ihren entsprechenden cyanobakteriellen Homologen, Rückschlüsse auf die Positionierung eines abgeleiteten Plastidenvorfahren innerhalb der cyanobakteriellen Phylogenie zu erhalten, scheiterte mit großer Wahrscheinlichkeit an Artefakten in der phylogenetischen Rekonstruktion. Diese wiesen typische Fehlerquellen für bekannte phylogenetische Probleme, wie einen signifikant erhöhten Anteil einzigartiger Austausche (Stiller und Hall, 1999) in den Aminosäuresequenzen der photosynthetischen Eukaryoten im Vergleich mit den cyanobakteriellen Sequenzen sowie um ein Vielfaches erhöhte Astlängen auf. Aus diesem Grund wurde hier der Fokus auf die weit größere Menge an Kerngenen cyanobakteriellen Ursprungs erweitert und Muster des Vorhandenseins beziehungsweise der Abwesenheit von Genen sowie die Aminosäureidentität von 611 Proteinfamilien im Vergleich mit den cyanobakteriellen Genomen aus-

gewertet. Das Genrepertoire des hypothetischen Vorfahren der Plastiden wies die größte Ähnlichkeit zu heutigen filamentösen, Heterozysten-differenzierenden Cyanobakterien (Untergruppen IV und V) auf. Eine mögliche überhöhte Ableitung der Verwandtschaft der untersuchten Sequenzen allein durch eine ähnliche Aminosäurenzusammensetzung der Sequenzen konnte mit Hilfe einer Analyse der prinzipiellen Komponenten (engl. *principal component analysis*) ausgeschlossen werden.

7 Anhang

Die folgenden Abschnitte enthalten das Anhangsmaterial zu den in Kapitel 5 aufgeführten Publikationen. Einige sehr umfangreiche Tabellen sind hier nicht aufgeführt. Die jeweilige vollständige Version des Anhangsmaterials kann von den entsprechenden Internetseiten der wissenschaftlichen Zeitschriften heruntergeladen werden (*Genome Biology and Evolution* online[1], *Molecular Biology and Evolution* online[2]).

1 http://www.gbe.oxfordjournals.org/
2 http://www.mbe.oxfordjournals.org/

7.1 Anhang zu Dagan *et al.* (2010)

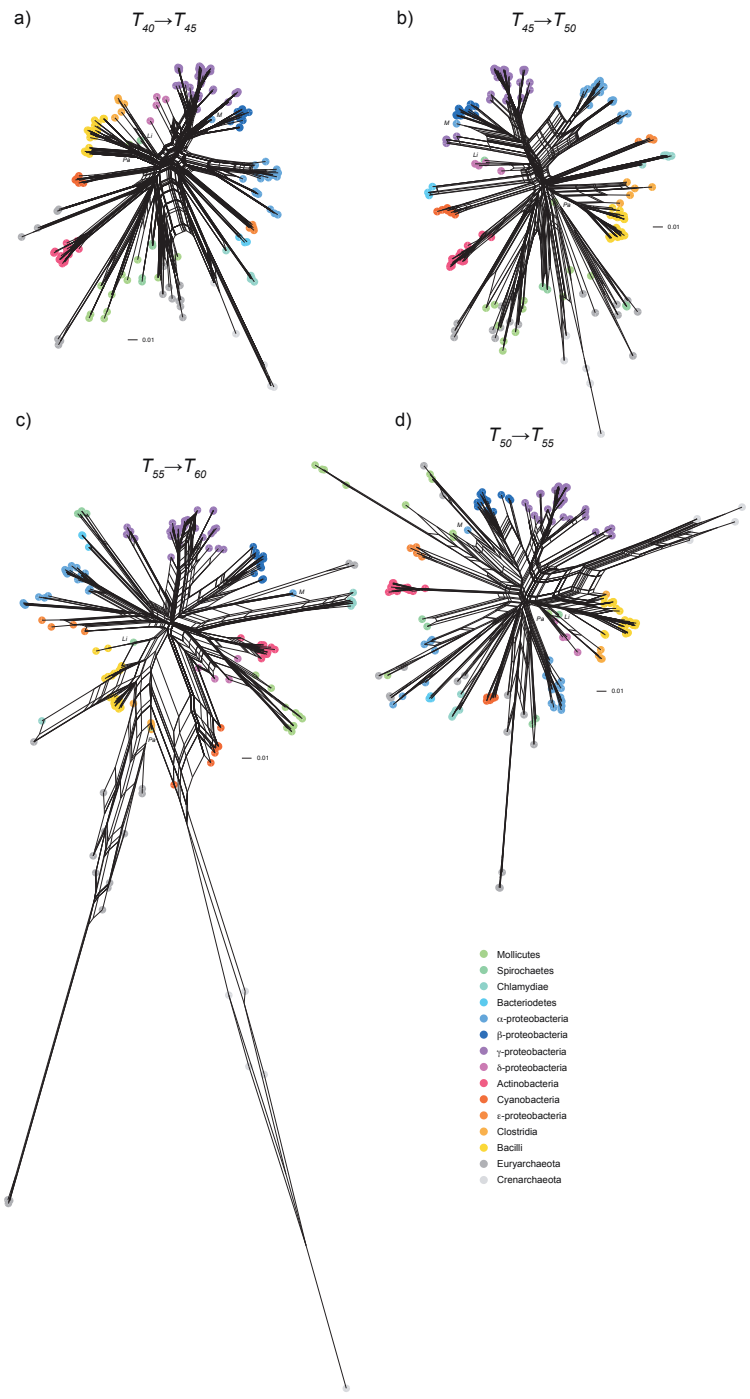


Figure S1: Protein family splits networks for higher protein similarity cutoffs.

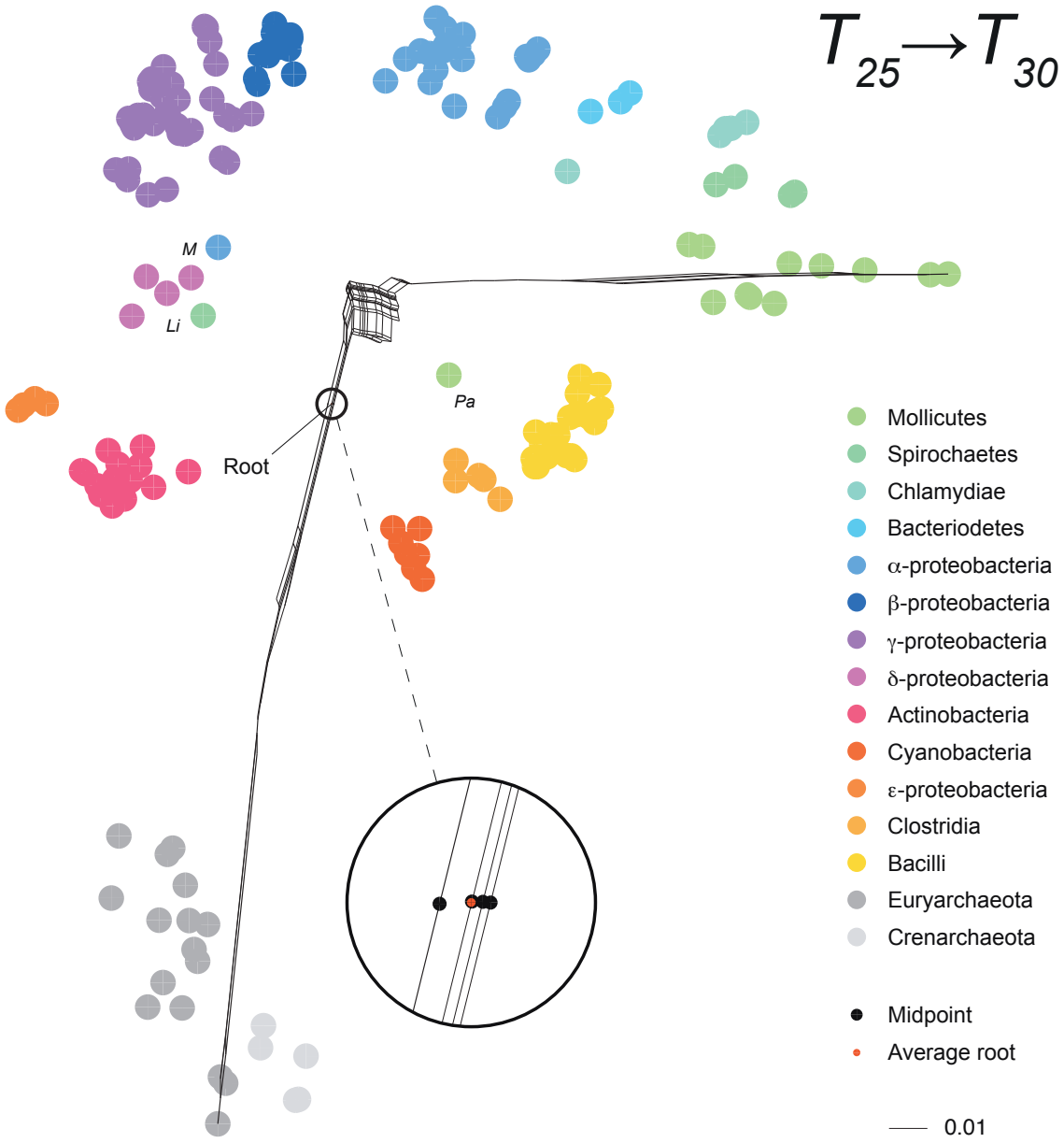


Figure S2: Root-pathways in the protein family splits network. Only splits separating the most distant pair of taxa are drawn. As an example for one possible minimum path between these taxa respective edges are highlighted (bold line). This path passes each split exactly one time.

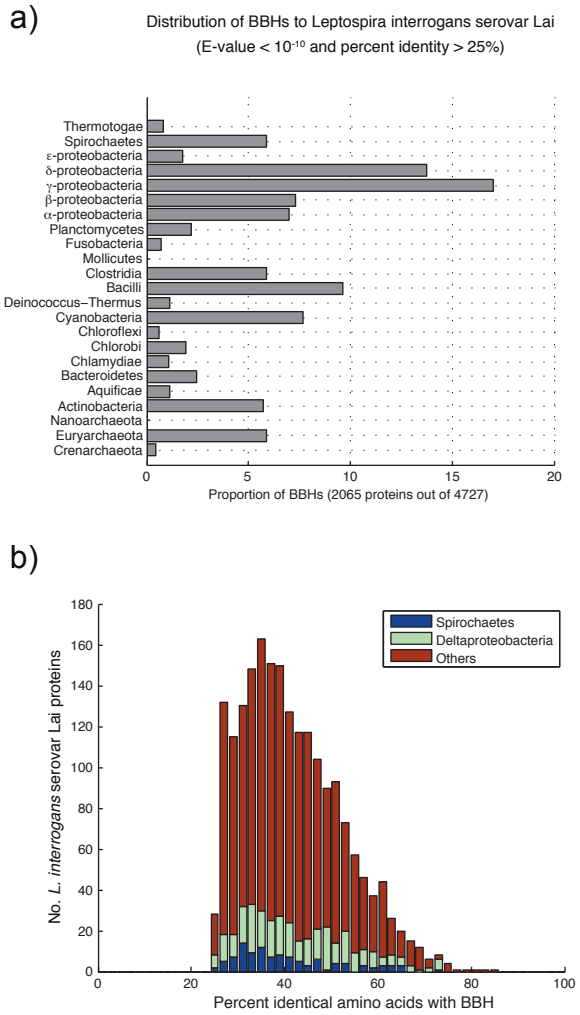


Figure S3: Nearest neighbors to *Leptospira interrogans* serovar Lai.

Table S1: Protein family sizes under different protein similarity thresholds.

	Total	Universal
T_{25}	53429	3831 (7%)
T_{30}	57743	3080 (5%)
T_{35}	62062	2352 (4%)
T_{40}	66205	1688 (3%)
T_{45}	69671	1121 (2%)
T_{50}	71616	676 (1%)
T_{55}	72011	367 (1%)
T_{60}	70959	172 (0.2%)

	Archaebacteria	Eubacteria
T_{25}	5101 (10%)	44497 (83%)
T_{30}	5831 (10%)	48832 (85%)
T_{35}	6664 (11%)	53046 (85%)
T_{40}	7293 (11%)	57224 (86%)
T_{45}	7778 (11%)	60772 (87%)
T_{50}	7922 (11%)	63018 (88%)
T_{55}	7672 (11%)	63972 (89%)
T_{60}	7062 (10%)	63725 (90%)

	Actinobacteria	Cyanobacteria	Proteobacteria
T_{25}	5268 (10%)	1471 (3%)	18152 (34%)
T_{30}	5977 (10%)	1676 (3%)	20918 (36%)
T_{35}	6710 (11%)	1872 (4%)	23727 (38%)
T_{40}	7430 (11%)	2021 (4%)	26913 (41%)
T_{45}	8099 (12%)	2055 (4%)	29951 (43%)
T_{50}	8516 (12%)	1953 (4%)	32382 (45%)
T_{55}	8743 (12%)	1719 (3%)	34134 (47%)
T_{60}	8751 (12%)	1373 (3%)	35022 (49%)

	α-proteo	β-proteo	γ-proteo
T_{25}	3277 (6%)	1920 (4%)	6221 (12%)
T_{30}	3946 (7%)	2327 (4%)	7316 (14%)
T_{35}	4632 (9%)	2825 (5%)	8540 (16%)
T_{40}	5479 (10%)	3445 (6%)	10109 (19%)
T_{45}	6310 (12%)	4252 (8%)	11754 (22%)
T_{50}	6923 (13%)	5053 (9%)	13327 (25%)
T_{55}	7401 (14%)	5988 (11%)	14595 (27%)
T_{60}	7624 (14%)	6816 (13%)	15680 (29%)

Table S2: Species that do not group with their taxonomical group.

Phytoplasma OY strain	The bacterium *Phytoplasma* OY strain (Phylum:Tenericutes, Class:Mollicutes, Candidatus *Phytoplasma asteris*) is an obligate pathogen inhabiting both plant and insect. As a result to its adaptation to the rich environment within the host, its genome has undergone a drastic reductive evolution that eliminated many metabolic pathways, far more than any other Mollicute (Oshima et al. 2004). In our splits networks the *Phytoplasma* OY strain never groups with the other Mollicutes (Figure 2), rather it clusters between Clostridia and Bacilli (phylum Firmicutes). Mollicutes were traditionally defined as Firmicutes and only recently were they separated into a new phylum (Ludwig et al. 2008). It is probably the reduced genome of *Phytoplasma* OY strain that leads to lack of Mollicute-specific gene families and hence its grouping with the more distantly related Firmicutes.
Magnetococcus sp. strain MC-1	The chemolithoautotrophic bacterium *Magnetococcus* sp. strain MC-1 (phylum: Proteobacteria, class unclassified) belongs to a group of proteobacterial magnetotactic cocci that are distantly related with the rest of magnetotactic bacteria (Schübbe et al. 2009). By the nearest neighbor measure, the genome of strain MC-1 is a mosaic of proteobacterial genes, where α-proteobacterial genes are the most frequent homologs (Esser et al. 2007). In our splits network the *Magnetococcus* sp. splits always within Proteobacteria but not with one of the subclasses in particular (Figure 2).
Leptospira interrogans serovar Lai	*Leptospira interrogans* serovar Lai (phylum: Spirochaetes) is a representative virulent of the Leptospirosis disease. This bacterium has a large genome in comparison to the other Spirochaetes and its physiology is quite different as well (Ren et al. 2003). In our split networks, *L. interrogans* serovar Lai never splits with the other Spirochaetes, and most commonly splits within the δ-Proteobacteria (Figure 2). An examination of the closest homologs to the proteins of *L. interrogans* serovar Lai by the nearest neighbor measure reveals a majority of proteins from δ- and γ- proteobacteria. Yet, sequence identity distribution of these homologs is not significantly higher in comparison to the background similarities in the whole dataset, suggesting that the high frequency of δ-proteobacterial nearest neighbors is due to small sample of completely sequenced Spirochaete genomes rather than frequent lateral gene transfer between the two groups (Figure S3).

Table S3: Iterative root inference within the network by the Jackknife approach. In each line the table shows the pair of species with maximum split weight distance. The following line shows the result after exclusion of this pair and re-rooting. In each step the location of the root is described.

No. Exclusions	Maximum split weight distance between species		Split weight distance	Root placed between	
0	*Thermoplasma acidophilum*	*Mycoplasma pneumoniae*	0.381	archaebacteria	eubacteria
1	*Sulfolobus acidocaldarius DSM 639*	*Mycoplasma genitalium*	0.370	archaebacteria	eubacteria
2	*Sulfolobus tokodaii*	*Mycoplasma gallisepticum*	0.351	archaebacteria	eubacteria
3	*Picrophilus torridus DSM 9790*	*Mycoplasma penetrans*	0.334	archaebacteria	eubacteria
4	*Thermoplasma volcanium*	*Borrelia garinii PBi*	0.329	archaebacteria	eubacteria
5	*Pyrobaculum aerophilum*	*Borrelia burgdorferi*	0.324	archaebacteria	eubacteria
6	*Sulfolobus solfataricus*	*Chlamydia muridarum*	0.317	archaebacteria	eubacteria
7	*Aeropyrum pernix*	*Ureaplasma urealyticum*	0.308	archaebacteria	eubacteria
8	*Pyrococcus horikoshii*	*Chlamydophila caviae*	0.303	archaebacteria	eubacteria
9	*Methanopyrus kandleri*	*Chlamydophila pneumoniae TW 183*	0.301	archaebacteria	eubacteria
10	*Methanococcus jannaschii*	*Chlamydia trachomatis*	0.289	archaebacteria	eubacteria
11	*Pyrococcus abyssi*	*Chlamydophila abortus S26 3*	0.285	archaebacteria	eubacteria
12	*Thermococcus kodakaraensis KOD1*	*Treponema pallidum*	0.280	archaebacteria	eubacteria
13	*Halobacterium sp*	*Treponema denticola ATCC 35405*	0.277	archaebacteria	eubacteria
14	*Archaeoglobus fulgidus*	*Mycoplasma hyopneumoniae 232*	0.279	archaebacteria	eubacteria
15	*Pyrococcus furiosus*	*Mycoplasma mobile 163K*	0.266	archaebacteria	eubacteria
16	*Methanococcus maripaludis S2*	*Mycoplasma pulmonis*	0.265	archaebacteria	eubacteria
17	*Methanobacterium thermoautotrophicum*	*Wolinella succinogenes*	0.258	archaebacteria	eubacteria
18	*Haloarcula marismortui ATCC 43049*	*Helicobacter hepaticus*	0.245	archaebacteria	eubacteria
19	*Methanosarcina mazei*	*Mesoplasma florum L1*	0.246	archaebacteria	eubacteria
20	*Methanosarcina acetivorans*	*Mycobacterium bovis*	0.242	archaebacteria	eubacteria
21	*Streptomyces avermitilis*	*Helicobacter pylori 26695*	0.178	Actinobacteria	other eubacteria

Table S4: Quartet tree topologies for ancient paralogs. Only comparisons of archaebacteria and other taxa are presented. The full table is found in the supplementary table (Table S5).

			Trees			t_{vert} trees grouped into rate models											
										$r_{lineage}$							
				t_{vert}		r_{global}				Group A faster			Group B faster			$r_{unknown}$	
Group A	**Group B**	**Ancient paralog pair**	No.	No.	Percent	No.	Percent of t_{vert}	No.	Percent of t_{vert}	No.	Percent of t_{vert}	Percent of $r_{lineage}$	No.	Percent of t_{vert}	Percent of $r_{lineage}$	No.	Percent of t_{vert}
Archaebacteria	Actinobacteria	total	1074	1052	97,95	638	60,65	364	34,60	36	3,42	9,89	328	31,18	90,11	50	4,75
Archaebacteria	Aquificae	total	67	67	100,00	45	67,16	18	26,87	9	13,43	50,00	9	13,43	50,00	4	5,97
Archaebacteria	Bacilli	total	1586	1573	99,18	886	56,33	603	38,33	177	11,25	29,35	426	27,08	70,65	84	5,34
Archaebacteria	Chlamydiae	total	330	318	96,36	153	48,11	146	45,91	13	4,09	8,90	133	41,82	91,10	19	5,97
Archaebacteria	Chlorobi	total	75	75	100,00	51	68,00	20	26,67	2	2,67	10,00	18	24,00	90,00	4	5,33
Archaebacteria	Chloroflexi	total	52	52	100,00	36	69,23	13	25,00	0	0,00	0,00	13	25,00	100,00	3	5,77
Archaebacteria	Clostridia	total	274	272	99,27	177	65,07	71	26,10	32	11,76	45,07	39	14,34	54,93	24	8,82
Archaebacteria	Cyanobacteria	total	424	390	91,98	211	54,10	136	34,87	20	5,13	14,71	116	29,74	85,29	43	11,03
Archaebacteria	Deinococcus-Thermus	total	159	159	100,00	108	67,92	42	26,42	5	3,14	11,90	37	23,27	88,10	9	5,66
Archaebacteria	Fusobacteria	total	52	52	100,00	34	65,38	16	30,77	4	7,69	25,00	12	23,08	75,00	2	3,85
Archaebacteria	Mollicutes	total	531	500	94,16	252	50,40	203	40,60	26	5,20	12,81	177	35,40	87,19	45	9,00
Archaebacteria	Planctomycetes	total	67	67	100,00	29	43,28	36	53,73	0	0,00	0,00	36	53,73	100,00	2	2,99
Archaebacteria	Spirochaetes	total	305	301	98,69	159	52,82	96	31,89	11	3,65	11,46	85	28,24	88,54	46	15,28
Archaebacteria	Thermotogae	total	73	73	100,00	43	58,90	19	26,03	10	13,70	52,63	9	12,33	47,37	11	15,07
Archaebacteria	Alphaproteobacteria	total	1249	1201	96,16	733	61,03	387	32,22	16	1,33	4,13	371	30,89	95,87	81	6,74
Archaebacteria	Betaproteobacteria	total	841	833	99,05	501	60,14	263	31,57	14	1,68	5,32	249	29,89	94,68	69	8,28
Archaebacteria	Deltaproteobacteria	total	244	241	98,77	139	57,68	75	31,12	15	6,22	20,00	60	24,90	80,00	27	11,20
Archaebacteria	Epsilonproteobacteria	total	230	223	96,96	137	61,43	67	30,04	8	3,59	11,94	59	26,46	88,06	19	8,52
Archaebacteria	Gammaproteobacteria	total	2293	2224	96,99	1279	57,51	787	35,39	77	3,46	9,78	710	31,92	90,22	158	7,10
		Summary	9926	9673	97,45	5611	58,01	3362	34,76	475	4,91	14,13	2887	29,85	85,87	700	7,24

Table S5: Quartet tree topologies reconstructed for ancient paralogs. Only comparisons of several different taxon groups with archaebacteria are presented. Comparisons are shown for all ancient paralogs separately.

Diese Tabelle ist im Anhangsmaterial der Publikation auf der Internetseite der wissenschaftlichen Zeitschrift Genome Biology and Evolution online[1] zu finden.

7.2 Anhang zu Roettger, Martin und Dagan (2009)

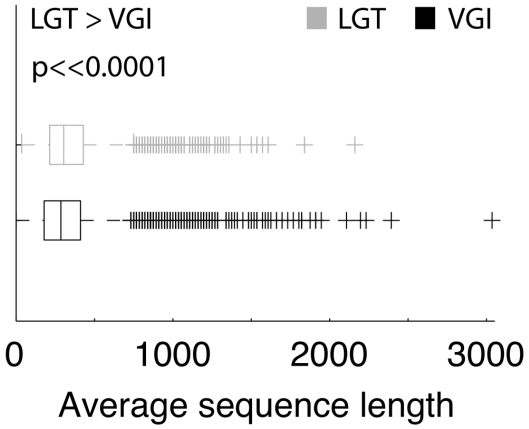


Figure S1: Distribution of mean sequence length per cluster for LGT and VGI. Differences in the distributions of the two groups were tested by the Wilcoxon non-parametric test (p-value presented at the top of graph).

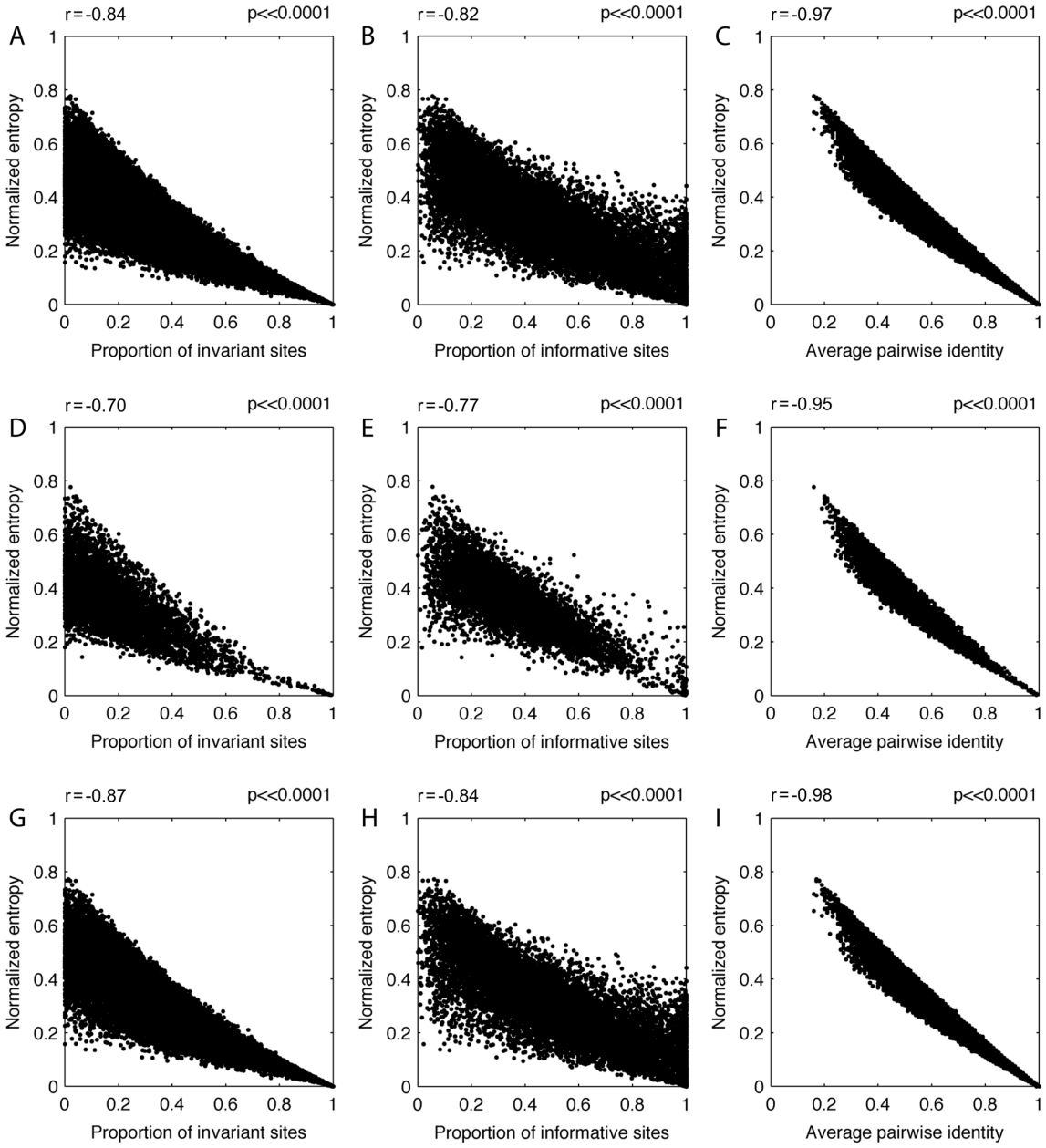


Figure S2: Correlation between proportion of invariant sites and normalized entropy, between proportion of informative sites and normalized entropy and between average pairwise identity and normalized entropy for the whole dataset (A), (B), (C) for the LGT alignments (D), (E), (F) and for the VGI alignments (G), (H), (I). Correlattion coefficient r and p-value are shown above each diagram.

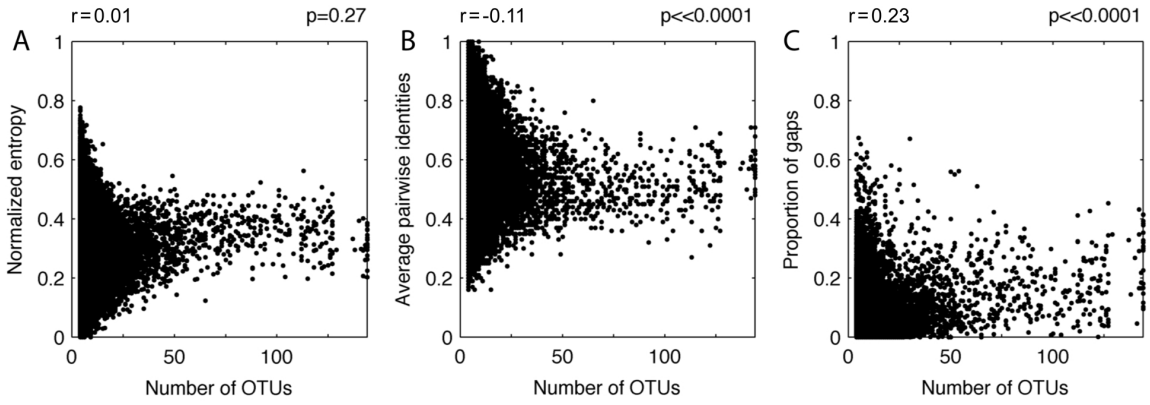


Figure S3: Correlation between number of OTUs and normalized entropy (A), average pairwise identity (B) and proportion of gaps (C). Correlation coefficient r and p-value are shown above each diagram.

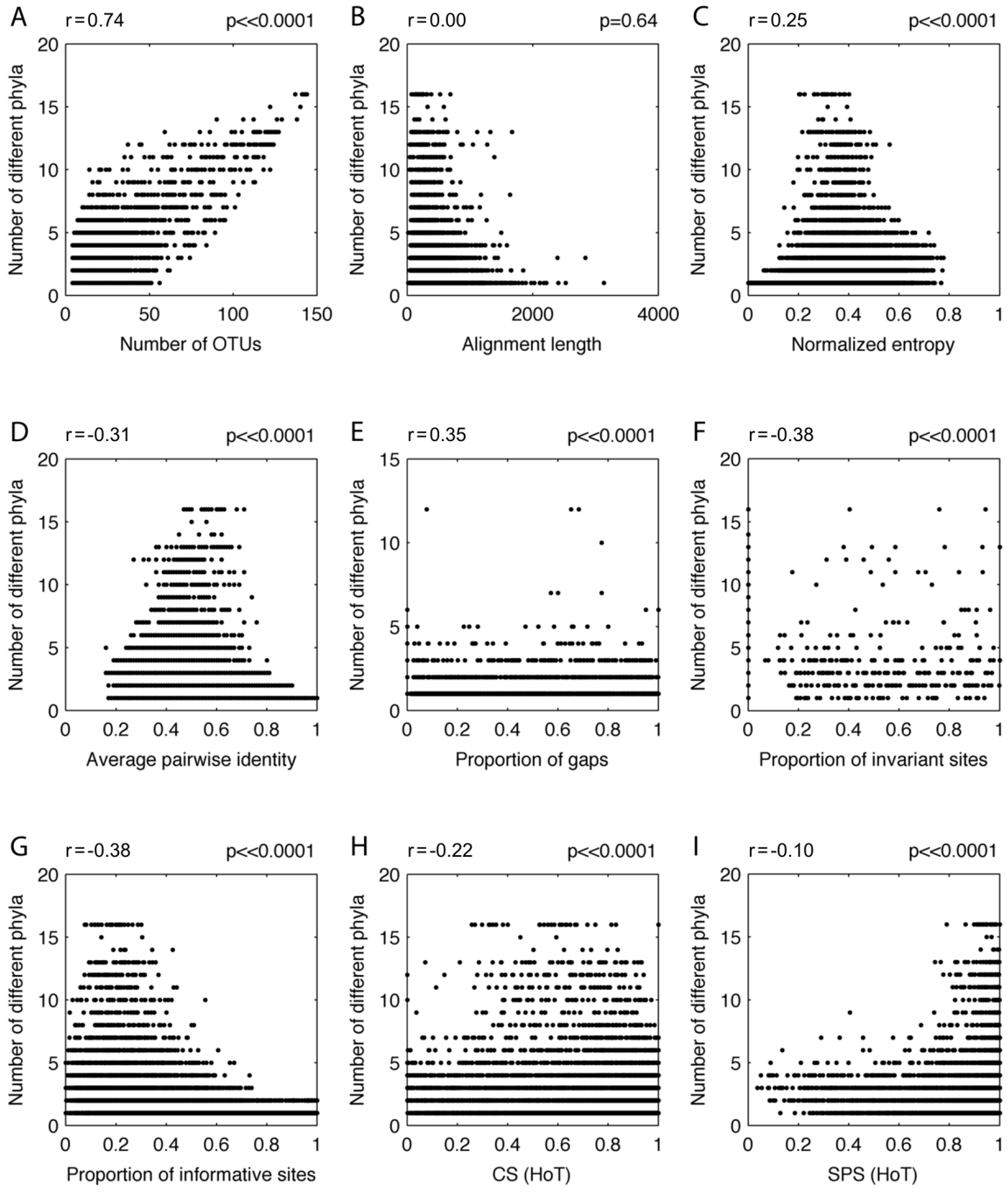


Figure S4: Correlation between the number of different phyla in the alignments and several different alignment features. Correlation coefficient r and p-value are shown above each diagram.

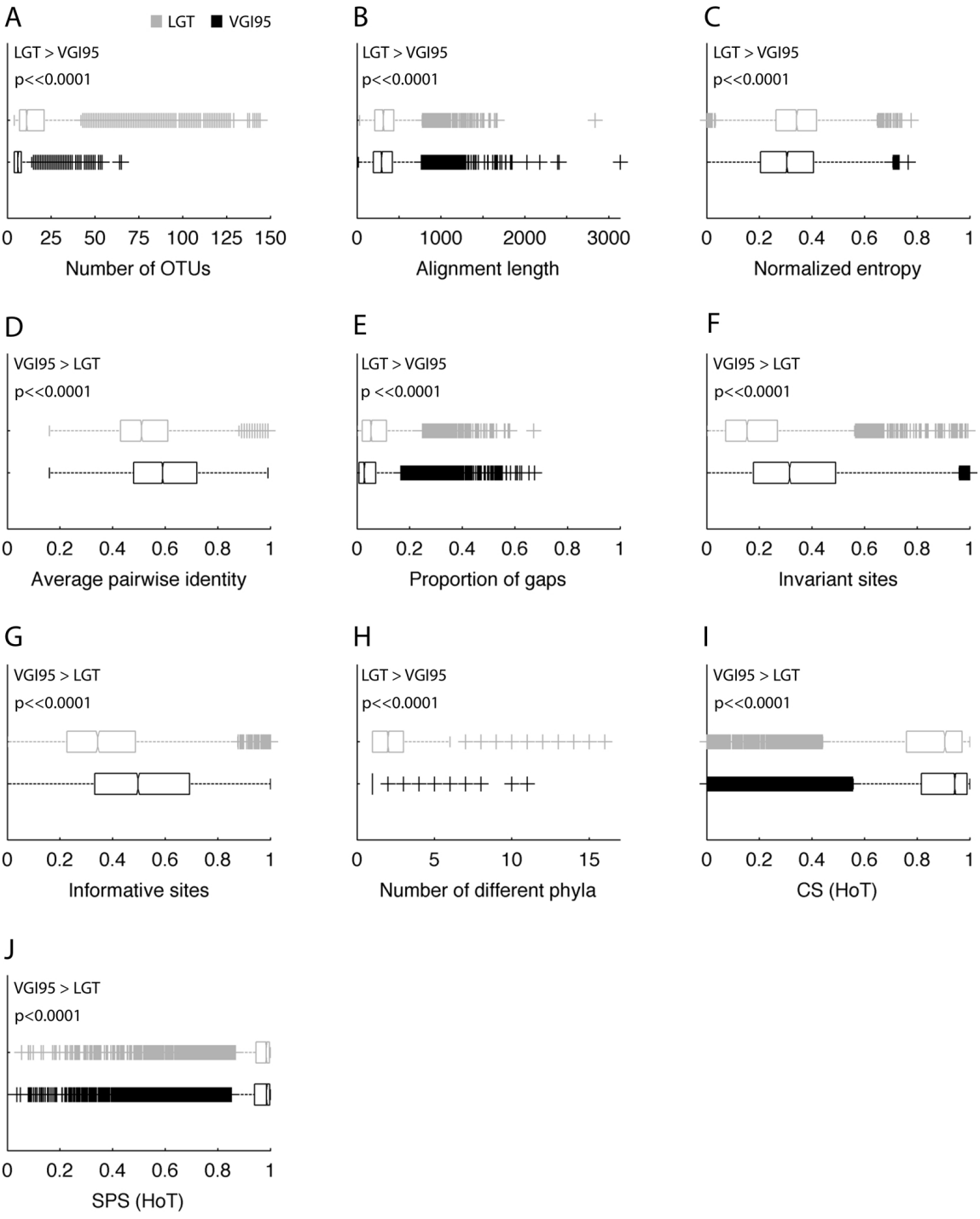


Figure S5: Distributions of alignment properties in the LGT and VGI95 groups (excluding unresolved trees). Differences in the distributions of the two groups were tested by the Wilcoxon non-parametric test (p-values presented at the top of each graph).

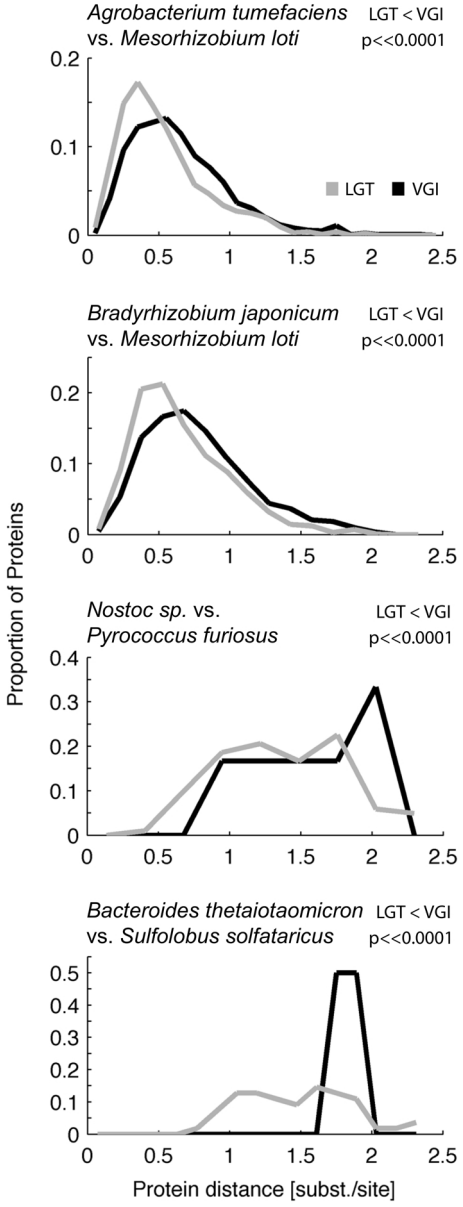


Figure S6: Comparison of protein pairwise distances between genome pairs found in LGT and VGI families. Distance distributions were compared using the Wilcoxon non-parametric test (p-values presented at the top right of the graph).

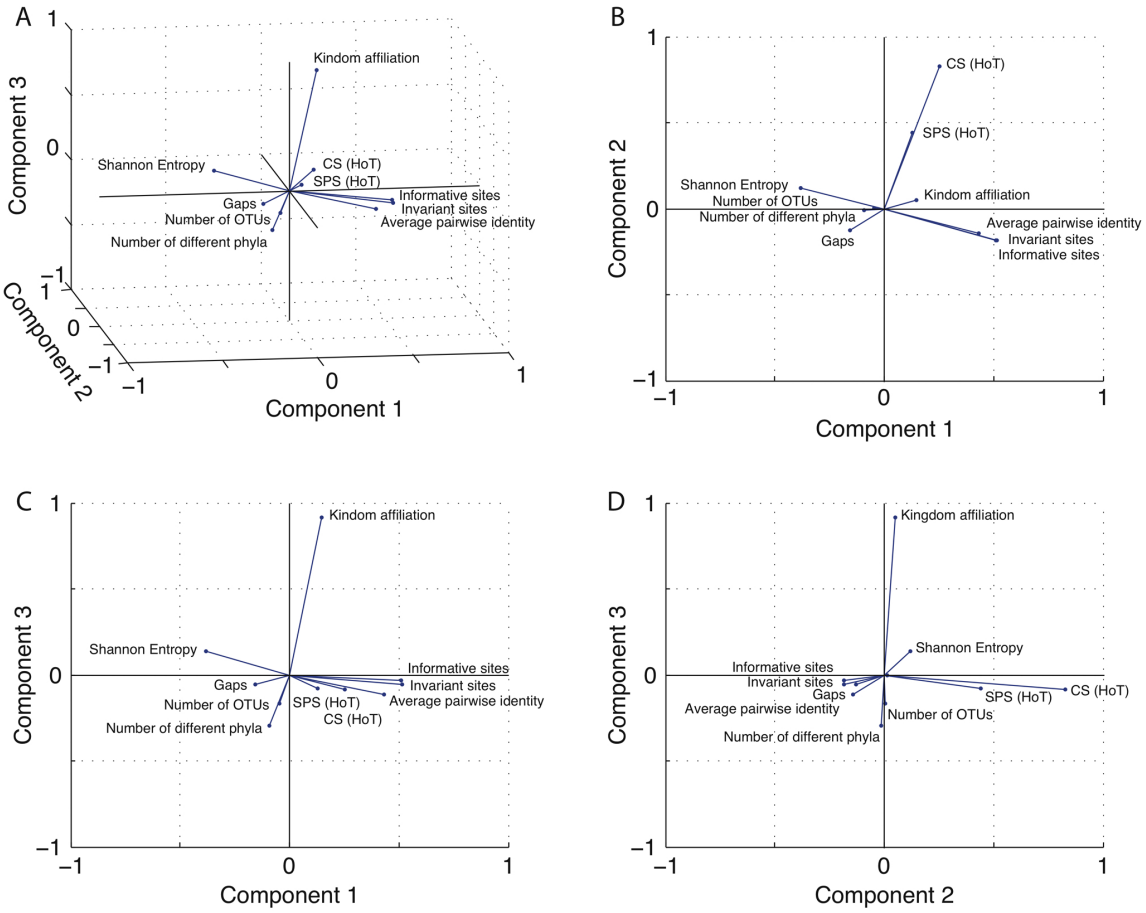


Figure S7: Principal component analysis: Contribution of alignment properties to the first three principal components. Axes represent principal components (factors), alignment properties are represented as vectors of their principal component coefficients. Three-dimensional view of the component space (A). Two dimensional views of every two respective components (B-C). The alignment length contribution to the first three principal components is only marginal, so that it was not plotted.

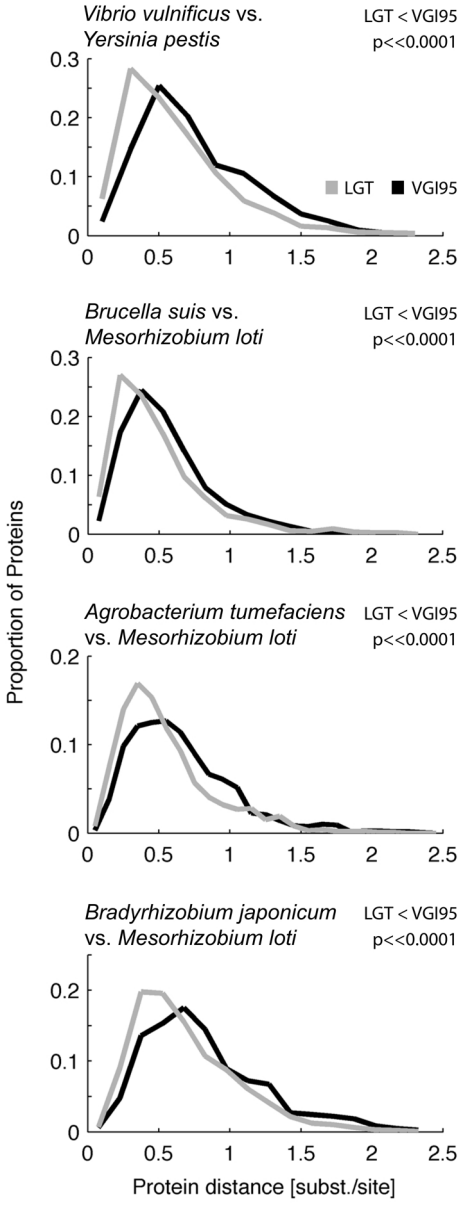


Figure S8: Comparison of protein pairwise distances between genome pairs found in LGT and VGI95 families (excluding alignments yielding unresolved trees). Distance distributions were compared using the Wilcoxon non-parametric test (p-values presented at the top right of the graph).

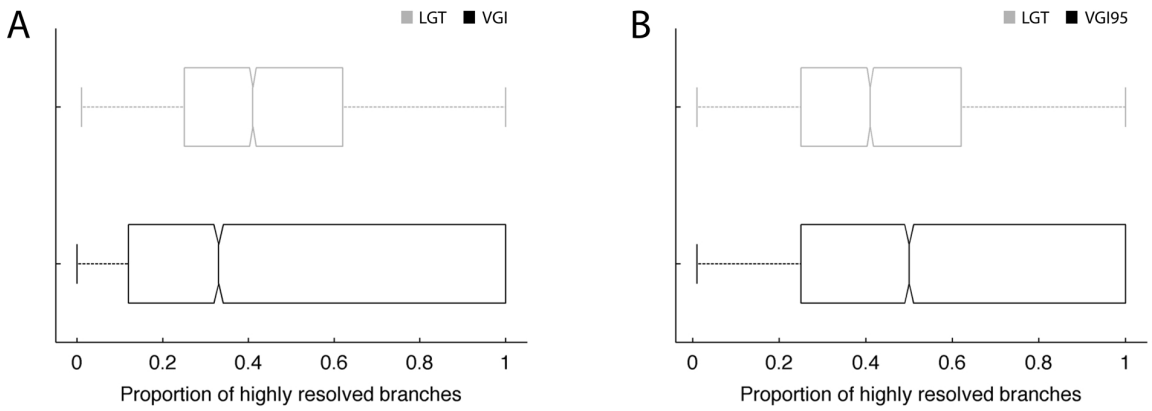


Figure S9: Distribution of highly resolved bipartitions for the LGT/VGI set (including unresolved trees) (A) and the LGT/VGI95 set (excluding unresolved trees) (B).

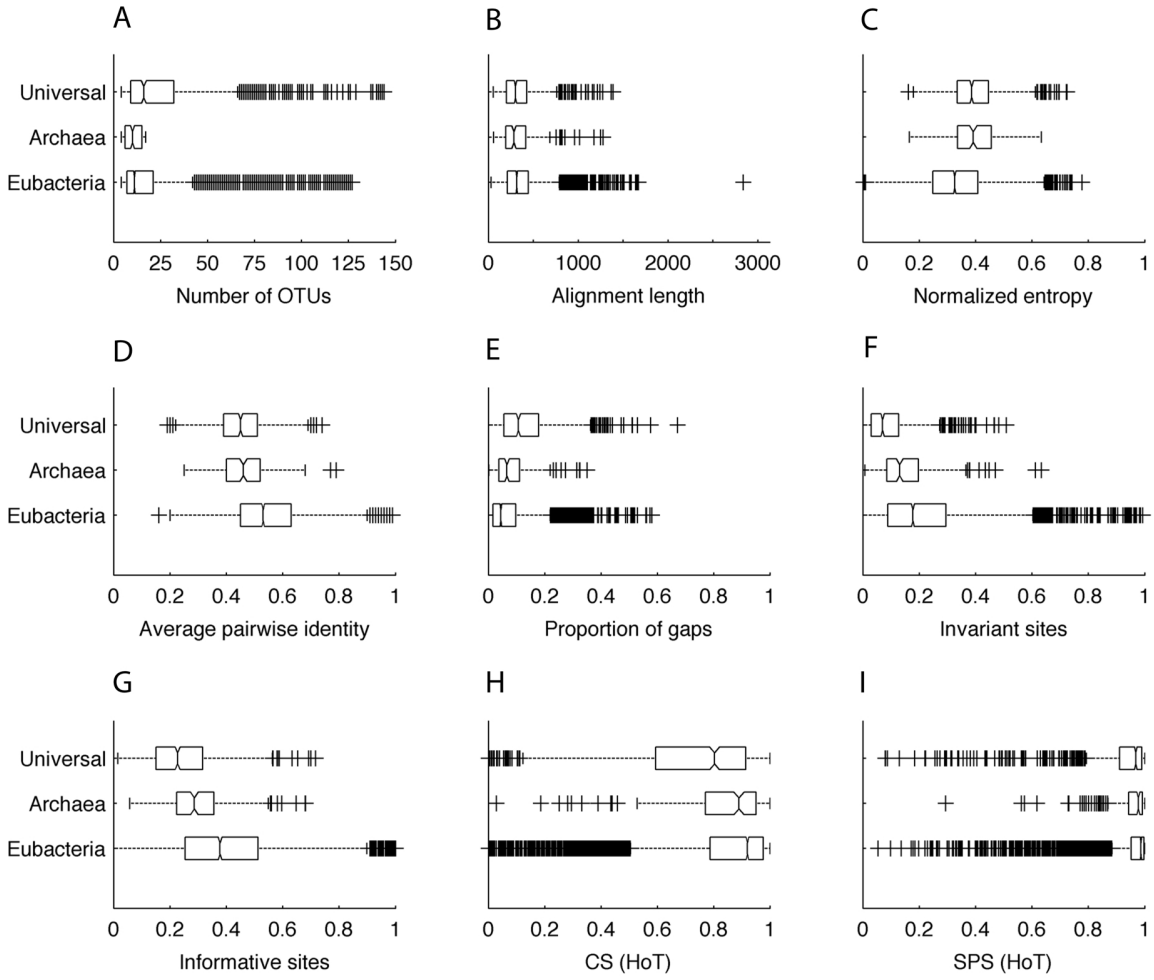


Figure S10: Differences in alignment features for alignments containing only eubacterial, only archaebacterial sequences, or sequences from both superkingdoms (universal) for LGT alignments only.

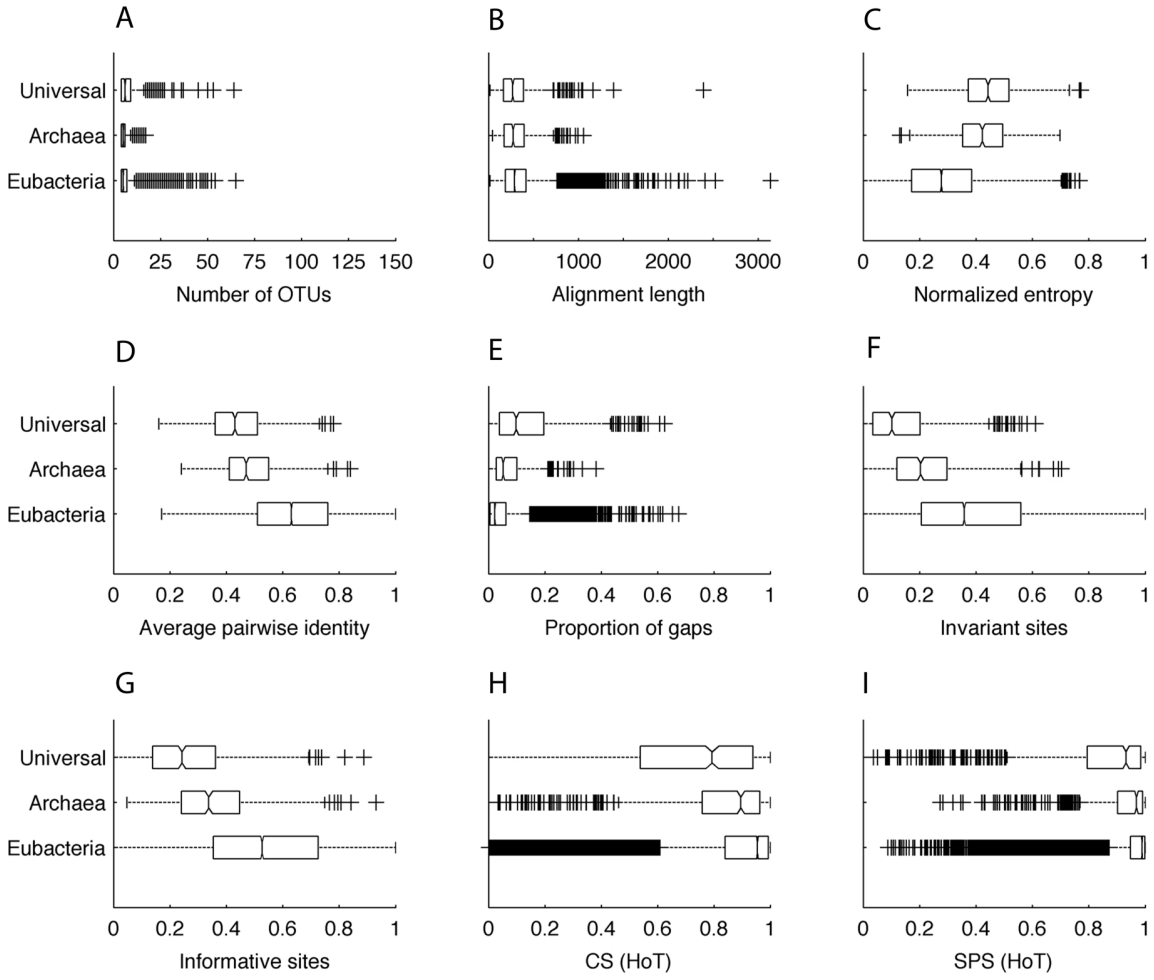


Figure S11: Differences in alignment features for alignments containing only eubacterial, only archaebacterial sequences, or sequences from both superkingdoms (universal) for VGI alignments only.

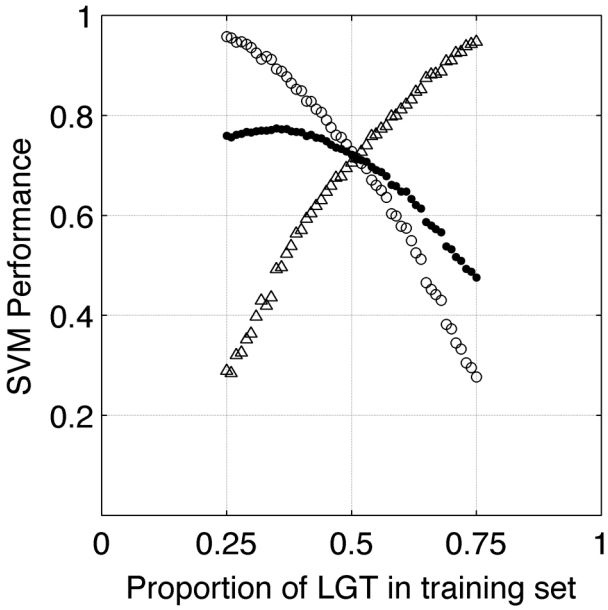


Figure S12: Performance of classification under different LGT proportions in the training set (excluding alignments yielding unresolved trees).

Table S1: Prediction of LGT/VGI using a support vector machine classifier trained with alignment properties. Alignment properties included in the training process are marked with ?. All possible combinations are shown. Definitions of accuracy, sensitivity and specificity can be found in the text.

Diese Tabelle ist im Anhangsmaterial der Publikation auf der Internetseite der wissenschaftlichen Zeitschrift Molecular Biology and Evolution online[2] zu finden.

Table S2: Prediction of LGT/VGI95 using a support vector machine classifier trained with alignment properties (excluding alignments yielding unresolved trees). Alignment properties included in the training process are marked with ?. All possible combinations are shown. Definitions of accuracy, sensitivity and specificity can be found in the text.

Diese Tabelle ist im Anhangsmaterial der Publikation auf der Internetseite der wissenschaftlichen Zeitschrift Molecular Biology and Evolution online[2] zu finden.

Table S3: Principal component coefficients for the 11 alignment properties for each factor.

	Component (factor)										
Alignment property	1	2	3	4	5	6	7	8	9	10	11
Number of OTUs	0.0463	-0.0023	0.1663	-0.5699	0.3106	-0.1318	-0.0680	0.0049	-0.016	0.7257	0.0044
Shannon entropy	0.3803	-0.1178	-0.1386	0.4017	0.2207	-0.0540	-0.4602	0.0135	0.1131	0.1739	0.5952
Average pairwise identity	-0.4338	0.1416	0.1138	-0.2610	-0.3777	-0.0310	0.0916	0.0076	-0.072	-0.045	0.7425
Proportion of gaps	0.1583	0.1282	0.0542	-0.2026	-0.1960	0.8681	-0.2659	0.154	0.1609	0.0381	-0.026
Proportion of invariant sites	-0.5155	0.1819	0.0582	0.2226	-0.2339	-0.1847	-0.6390	0.0251	0.1598	0.2068	-0.287
Proportion of informative sites	-0.5122	0.1860	0.0319	0.2095	0.7275	0.3167	0.1252	0.0146	0.0076	-0.052	0.0994
Alignment length	-0.0062	-0.0126	0.0031	-0.0304	-0.0195	0.1460	-0.0927	-0.983	-0.049	0.0075	0.0009
CS (HoT)	-0.2530	-0.8266	0.0823	0.0489	-0.0364	0.1910	-0.1233	0.0733	-0.425	0.0621	-0.019
SPS (HoT)	-0.1300	-0.4428	0.0795	-0.1318	0.0403	-0.0504	0.1007	-0.05	0.8559	-0.113	0.0327
Number of different phyla	0.0936	0.0116	0.2932	-0.4250	0.2861	-0.1737	-0.4767	0.0281	-0.119	-0.608	0.0069
Kingdom affiliation	-0.1488	-0.0485	-0.9131	-0.3315	0.0669	-0.0114	-0.1390	0.019	-0.011	-0.086	-0.01

Table S4: Variance coverage of the factors extracted by principal component analysis for the whole dataset and categorization of the main properties building the factor (description).

Component (Factor)	Variance (Eigenvalue)	Percentage of total variance	Cumulative	Description
1	0.1998	61.71	61.71	Variability
2	0.0463	14.30	76.01	Reliability (CS)
3	0.0316	9.76	85.77	Kingdom affiliation
4	0.0162	5.01	90.78	-
5	0.0091	2.81	93.60	Informative sites
6	0.0071	2.18	95.78	Gap content
7	0.0056	1.74	97.52	-
8	0.0043	1.32	98.84	Alignment length
9	0.0020	0.61	99.45	Reliability (SPS)
10	0.0015	0.46	99.91	Number of OTUs
11	0.0003	0.09	100.00	Variability

7.3 Anhang zu Dagan *et al.* (2012)

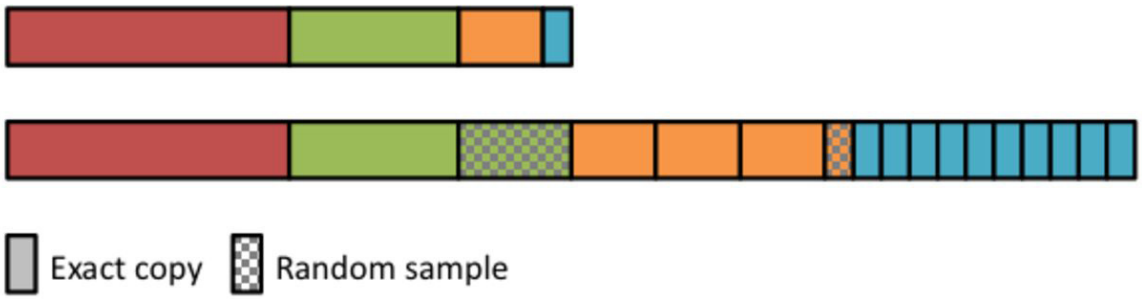


Figure S1: Producing a weighted concatenated sequence alignment. In the common alignment concatenation procedure (top bar), a few long genes may contribute the majority of alignment columns to the resulting alignment. In our data, for example, 39 % of the genes account for 61 % of the columns. In weighted concatenation (bottom bar), each gene is over-sampled to the size of the longest gene, by concatenating exact repeats of its alignment, and the remainder being a random sample of columns (chequered blocks). This ensures that all genes contribute the same number of columns to the concatenated alignment.

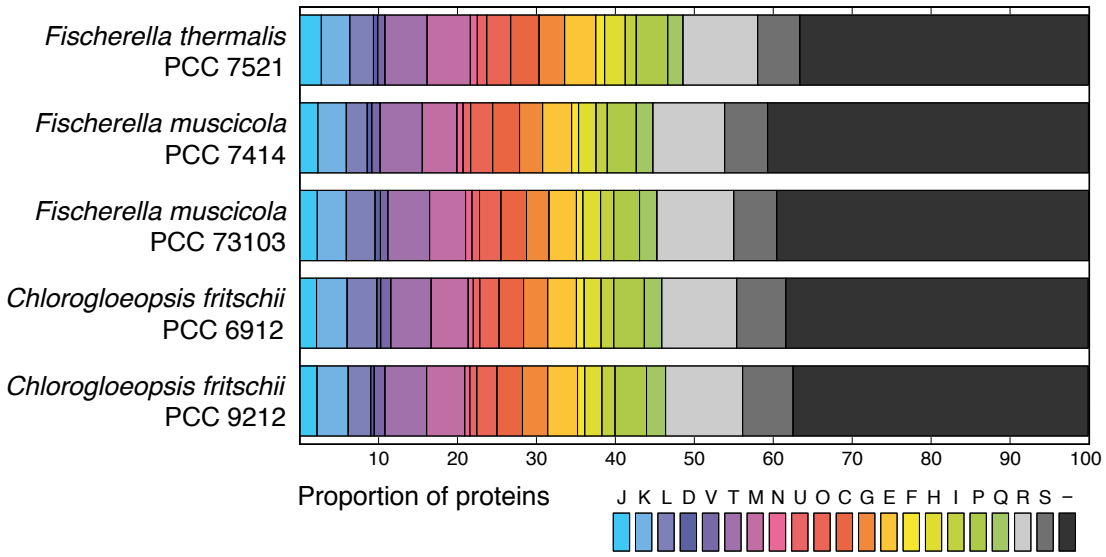


Figure S2: Functional classification of proteins in genomes of Subsection V cyanobacteria. CDSs in each sequenced cyanobacterium were classified into functional categories according to the COG functional classification (Tatusov *et al.*, 2001). Functional categories are colored according to the one letter code: J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; D, cell cycle control, cell division, chromosomal partitioning; V, defence mechanisms; T, signal transduction; M, cell wall/membrane/envelope biosynthesis; N, cell motility; Z, cytoskeleton; U, intracellular trafficking, secretion, vesicular transport; O, posttranslational modification, protein turnover, chaperones; C, energy production and conversion; transport and metabolism of: G, carbohydrate; E, amino acids; F, nucleotides; H, coenzymes; I, lipids; P, inorganic ions; Q, secondary metabolites. R, general function prediction only; S, function unknown; -, no COG annotation.

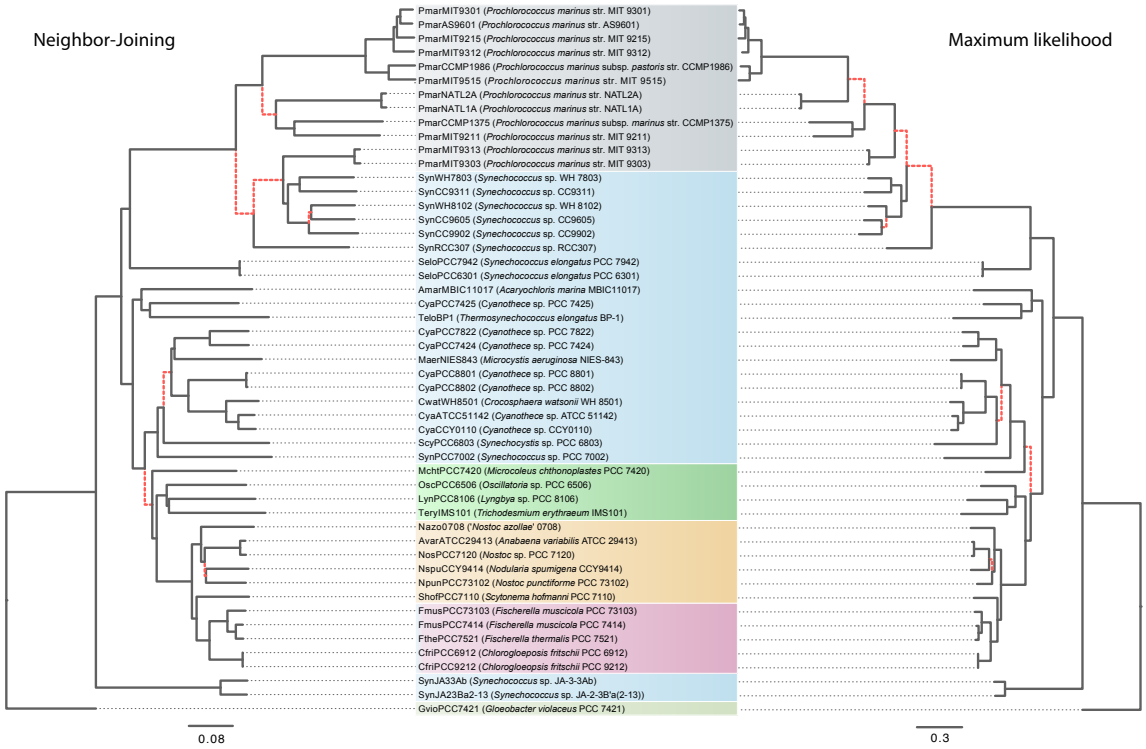


Figure S3: Maximum likelihood and neighbor-joining phylogenies reconstructed from the concatenated alignment of 324 universal cyanobacterial single-copy proteins. The trees are rooted by *Gloeobacter violaceus* PCC 7421. Bootstrap support on neighbor-joining tree was 100 % for all branches. Incompatible branches between neighbor-joining and maximum likelihood phylogeny highlighted as dashed red lines. Prochlorales (grey), Chroococcales (blue), Oscillatoriales (green), Nostocales (orange), Stigonematales (red), Gloeobacterales (light green).

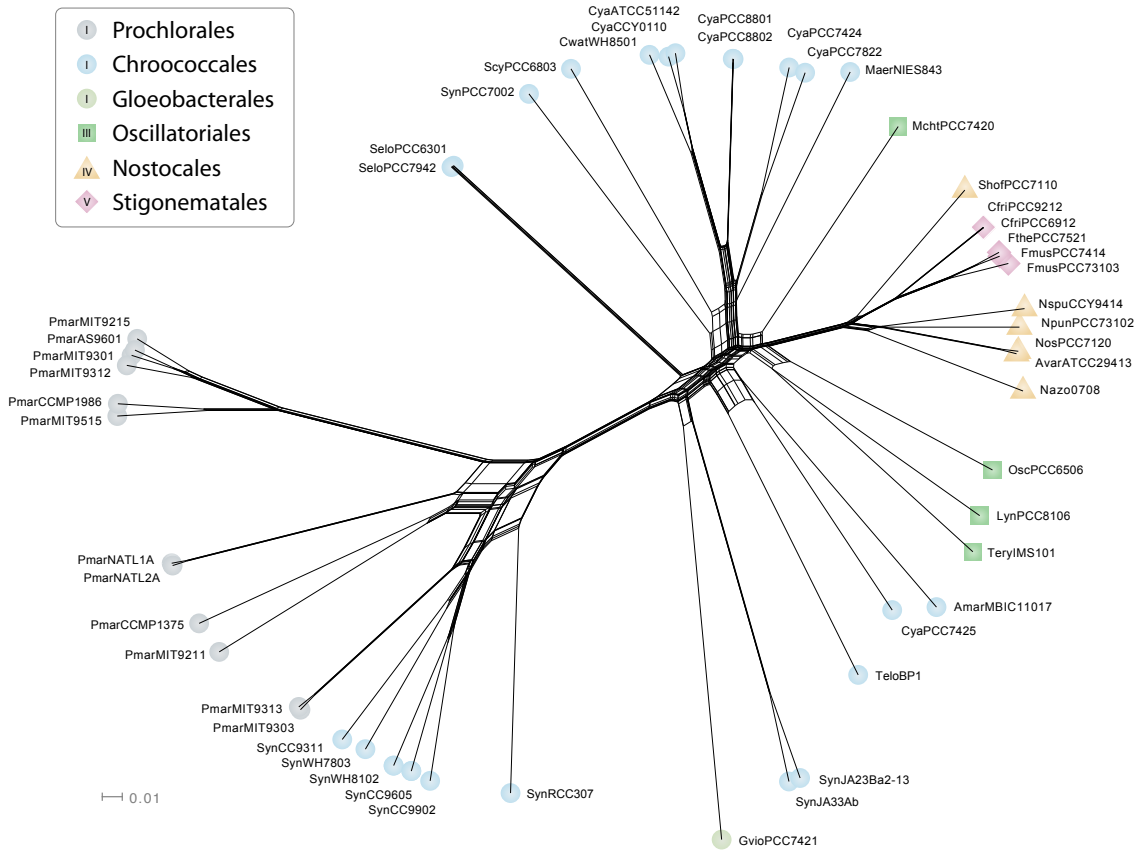
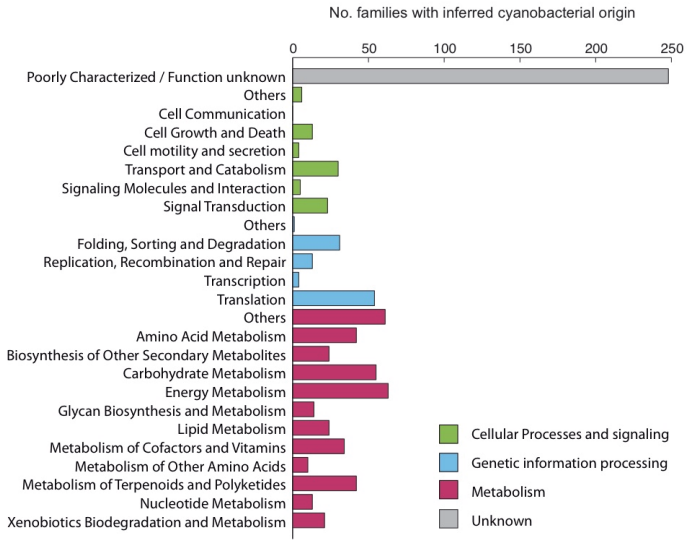


Figure S4: NeighborNet reconstructed from weighted concatenated alignment of 324 universal cyanobacterial single-copy proteins.



KEGG class	Main category	Subcategory
Unclassified; Poorly Characterized; General function prediction only	Poorly Characterized / Function unknown	-
Unclassified; Poorly Characterized; Function unknown		
Unclassified; Cellular Processes and Signaling; Electron transfer carriers	Cellular Processes and Signalling	Others
Unclassified; Cellular Processes and Signaling; Membrane and intracellular structural molecules		Others
Unclassified; Cellular Processes and Signaling; Pores ion channels		Others
Unclassified; Cellular Processes and Signaling; Sporulation		Others
Cellular Processes; Cell Communication		Cell Communication
Cellular Processes; Cell Growth and Death		Cell Growth and Death
Cellular Processes; Cell Motility		Cell motility and secretion
Unclassified; Cellular Processes and Signaling; Cell motility and secretion		Cell motility and secretion
Cellular Processes; Transport and Catabolism		Transport and Catabolism
Unclassified; Cellular Processes and Signaling; Other ion-coupled transporters		Transport and Catabolism
Unclassified; Cellular Processes and Signaling; Other transporters		Transport and Catabolism
Environmental Information Processing; Membrane Transport		Transport and Catabolism
Environmental Information Processing; Signaling Molecules and Interaction		Signaling Molecules and Interaction
Environmental Information Processing; Signal Transduction		Signal Transduction
Unclassified; Cellular Processes and Signaling; Signal transduction mechanisms		Signal Transduction
Unclassified; Genetic Information Processing; Others	Genetic Information Processing	Others
Genetic Information Processing; Folding, Sorting and Degradation		Folding, Sorting and Degradation
Unclassified; Genetic Information Processing; Protein folding and associated processing		Folding, Sorting and Degradation
Genetic Information Processing; Replication and Repair		Replication, Recombination and Repair
Unclassified; Genetic Information Processing; Replication, recombination and repair proteins		Replication, Recombination and Repair
Genetic Information Processing; Transcription		Transcription
Genetic Information Processing; Translation		Translation
Unclassified; Genetic Information Processing; Translation proteins		Translation
Unclassified; Metabolism; Others	Metabolism	Others
Metabolism; Enzyme Families		Others
Metabolism; Amino Acid Metabolism		Amino Acid Metabolism
Unclassified; Metabolism; Amino acid metabolism		Amino Acid Metabolism
Metabolism; Biosynthesis of Other Secondary Metabolites		Biosynthesis of Other Secondary Metabolites
Unclassified; Metabolism; Biosynthesis and biodegradation of secondary metabolites		Biosynthesis of Other Secondary Metabolites
Metabolism; Carbohydrate Metabolism		Carbohydrate Metabolism
Unclassified; Metabolism; Carbohydrate metabolism		Carbohydrate Metabolism
Metabolism; Energy Metabolism		Energy Metabolism
Unclassified; Metabolism; Energy metabolism		Energy Metabolism
Metabolism; Glycan Biosynthesis and Metabolism		Glycan Biosynthesis and Metabolism
Metabolism; Lipid Metabolism		Lipid Metabolism
Metabolism; Metabolism of Cofactors and Vitamins		Metabolism of Cofactors and Vitamins
Unclassified; Metabolism; Metabolism of cofactors and vitamins		Metabolism of Cofactors and Vitamins
Metabolism; Metabolism of Other Amino Acids		Metabolism of Other Amino Acids
Metabolism; Metabolism of Terpenoids and Polyketides		Metabolism of Terpenoids and Polyketides
Metabolism; Nucleotide Metabolism		Nucleotide Metabolism
Unclassified; Metabolism; Nucleotide metabolism		Nucleotide Metabolism
Metabolism; Xenobiotics Biodegradation and Metabolism		Xenobiotics Biodegradation and Metabolism

Figure S5: Functional categories of plant protein families of cyanobacterial origin. Functional categories of CDSs were inferred by protein sequence similarity search in KEGG database (Kanehisa *et al.*, 2012). Functional main categories were assigned according to the annotation of the best blast hit in the database with E-value $\leq 10^{-10}$ and identity of at least 25 %. Protein families were classified into the functional categories according to the most abundant functional assignment of their member proteins.

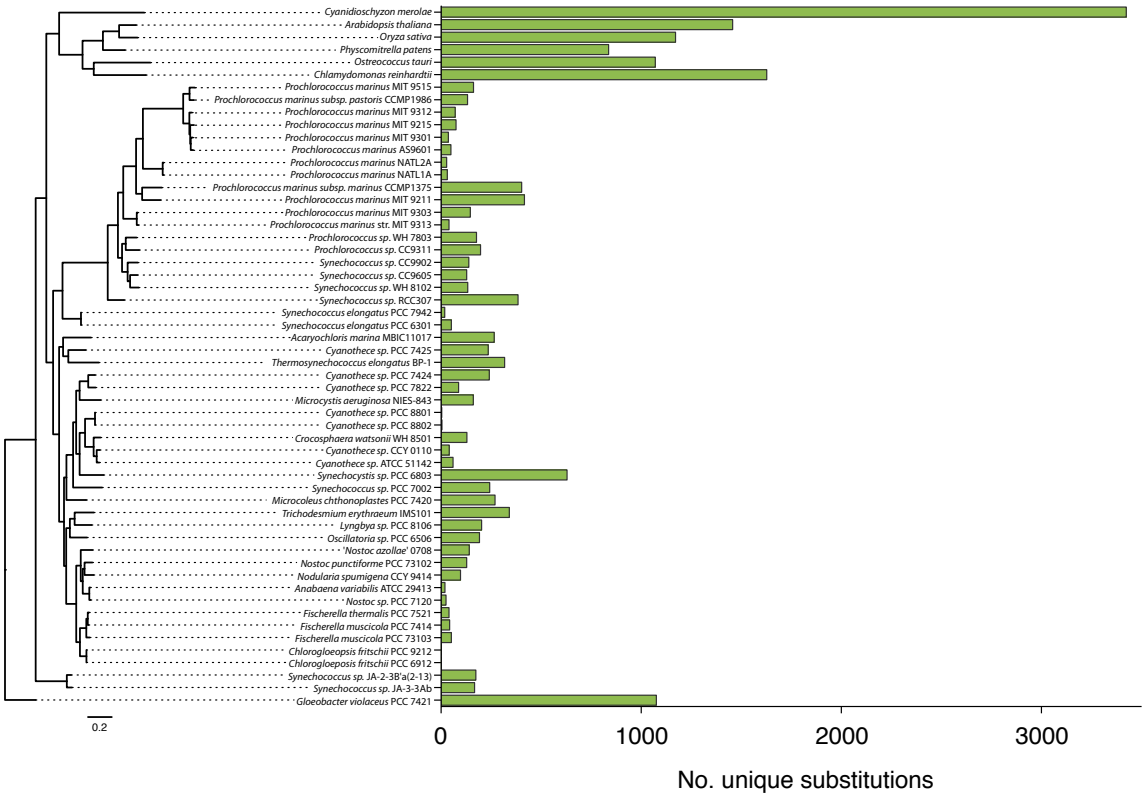


Figure S6: Topology and distribution of unique characters in a maximum likelihood phylogenetic tree reconstructed from universal cyanobacterial genes and plant nuclear genes of cyanobacterial origin. The frequency of unique substitutions per organism was calculated from the concatenated alignment of 23 universal protein families common to cyanobacteria and plants. Amino acid position is counted as unique if the character state at that position is unique to a particular organism. The tree is rooted by *Gloeobacter violaceus* PCC 7421.

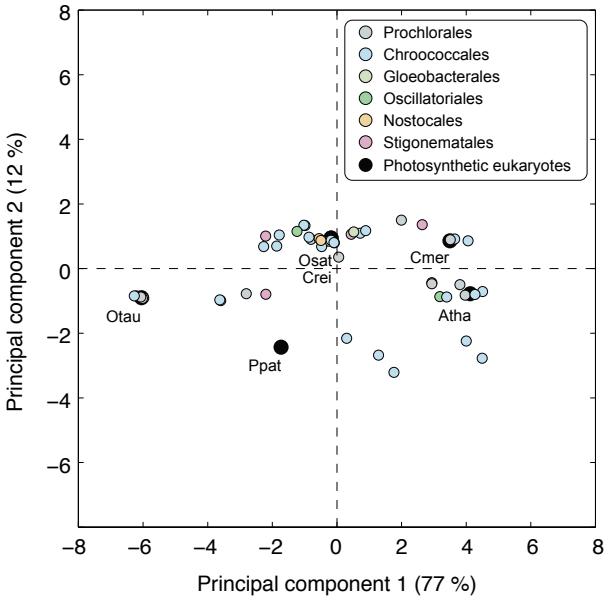


Figure S7: Principal component analysis of amino acid usage within 611 proteins of endosymbiotic origin. Average amino acid frequencies for respective amino acid sequences for each cyanobacterium or photosynthetic eukaryote were used to perform a principal component analysis using princomp function of MATLAB (The MathWorks Inc., 2012). The first two principal components are shown. Sequences of photosynthetic eukaryotes are indicated as black circles. Atha: *Arabidopsis thaliana*, Cmer: *Cyanidioschyzon merolae*, Crei: *Chlamydomonas reinhardtii*, Osat: *Oryza sativa*, Otau: *Ostreococcus tauri*, Ppat: *Physcomitrella patens*.

Table S1: Protein families identified as unique and common to all filamentous cyanobacteria (Subsections III, IV and V). Best BLAST hits in non-filamentous cyanobacteria, and outside cyanobacteria are also cited.

Diese Tabelle ist im Anhangsmaterial der Publikation auf der Internetseite der wissenschaftlichen Zeitschrift Genome Biology and Evolution online[1] zu finden.

Table S2: Protein families identified as specific for heterocyst forming cyanobacteria (Subsections IV and V). Best BLAST hits in non-heterocyst forming cyanobacteria, and outside cyanobacteria are also cited.

Diese Tabelle ist im Anhangsmaterial der Publikation auf der Internetseite der wissenschaftlichen Zeitschrift Genome Biology and Evolution online[1] zu finden.

Table S3: Protein families identified as uniquely shared among heterocystous cyanobacteria of Subsection V. Best BLAST hits in other cyanobacteria, and outside cyanobacteria are also cited.

Diese Tabelle ist im Anhangsmaterial der Publikation auf der Internetseite der wissenschaftlichen Zeitschrift Genome Biology and Evolution online[1] zu finden.

Literaturverzeichnis

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 21:2104–2105.

Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC. 1992. Sequence identification of 2,375 human brain genes. Nature. 355:632–634.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 252:1651–1656.

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. 5:e1000262.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res. 25:3389–3402.

Bandelt HJ, Dress AWM. 1992. A canonical decomposition theory for metrics on a finite set. Adv Math. 92:47–105.

Barbançois CH. 1816. Observation sur la filiation des animaux, depuis le polype jusqu'au singe. J Phys Chim Hist Nat Arts. 82:444–448.

Bryant D, Moulton V. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 21:255–265.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Cavalier-Smith T. 2006. Cell evolution and earth history: stasis and revolution. Philos Trans R Soc Lond B Biol Sci. 361:969–1006.

Cavalier-Smith T. 2009. Predation and eukaryote cell origins: a coevolutionary perspective. Int J Biochem Cell Biol. 41:307–322.

Cavalier-Smith T. 2010a. Deep phylogeny, ancestral groups and the four ages of life. Philos Trans R Soc Lond B Biol Sci. 365:111–132.

Cavalier-Smith T. 2010b. Origin of the cell nucleus, mitosis and sex: roles of intracellular coevolution. Biol Direct. 5:7.

Cho JC, Tiedje JM. 2001. Bacterial species determination from DNA-DNA hybridization by using genome fragments and DNA microarrays. Appl Environ Microbiol. 67:3677–3682.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311:1283–1287.

Coordinators NR. 2013. Database resources of the National Center for Biotechnology Information. Nucl Acids Res. 41:D8–D20.

Criscuolo A, Gribaldo S. 2011. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. Mol Biol Evol. 28:3019–3032.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105:10039–10044.

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

Dagan T, Roettger M, Bryant D, Martin W. 2010. Genome networks root the tree of life between prokaryotic domains. Genome Biol Evol. 2:379–392.

Dagan T, Roettger M, Stucken K, *et al*. (17 co-authors). 2012. Genomes of stigonematalean cyanobacteria (Subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome Biol Evol. 5:31–44.

Darwin C. 1859. On the origin of species. London: John Murray.

de Duve C. 2007. The origin of eukaryotes: a reappraisal. Nat Rev Genet. 8:395–403.

Deschamps P, Colleoni C, Nakamura Y, *et al*. (17 co-authors). 2008. Metabolic symbiosis and the birth of the plant kingdom. Mol Biol Evol. 25:536–548.

Dessimoz C, Gil M. 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11:R37.

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 25:748–761.

Doolittle WF. 1999. Phylogenetic classification and the universal tree. Science. 284:2124–2128.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. 32:1792–1797.

Enright AJ, van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucl Acids Res. 30:1575–1584.

Esser C, Martin W, Dagan T. 2007. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 3:180–184.

Faith DP. 1985. Distance methods and the approximation of most-parsimonious trees. Syst Zool. 34:312–325.

Farris JS. 1972. Estimating phylogenetic trees from distance matrices. American Naturalist. 106:645–668.

Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool. 22:240–249.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27:401–410.

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

Feng DF, Doolittle RF. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol. 25:351–360.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool. 19:99–113.

Fitch WM. 1981. A non-sequential method for constructing trees and hierarchical classifications. J Mol Evol. 18:30–37.

Fitch WM. 1995. Use for evolutionary trees. Philos Trans R Soc Lond B Biol Sci. 349:93–102.

Fitch WM. 1997. Networks and viral evolution. J Mol Evol. 44 Suppl 1:S65–S75.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 269:496–512.

Gotoh O. 1996. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol. 264:823–838.

Grasso C, Lee C. 2004. Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics. 20:1546–1556.

Graur D, Li WH. 2000. Fundamentals of molecular evolution. Sunderland, Massachusetts: Sinaur Associates, Inc.

Grünheit N. 2010. Clusteranalyse und Phylogenomik eukaryotischer Gene prokaryotischen Ursprungs. Ph.D. thesis, Heinrich-Heine-Universität, Düsseldorf.

Haeckel E. 1866. Generelle morphologie der organismen. Berlin: G. Reimer.

Haeckel E. 1874. Anthropogenie. Leipzig: Engelmann.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 89:10915–10919.

Higgins DG, Blackshields G, Wallace IM. 2005. Mind the gaps: progress in progressive alignment. Proc Natl Acad Sci U S A. 102:10411–10412.

Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet. 4:275–284.

Holland B, Delsuc F, Moulton V. 2005. Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods. Syst Biol. 54:66–76.

Hubbard TJP, Aken BL, Beal K, *et al*. (57 co-authors). 2007. Ensembl 2007. Nucl Acids Res. 35:D610–D617.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294:2310–2314.

Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. Nucl Acids Res. 36:D250–D254.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. Nucl Acids Res. 40:D109–D114.

Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucl Acids Res. 30:3059–3066.

Kawakita A, Sota T, Ascher JS, Ito M, Tanaka H, Kato M. 2003. Evolution and phylogenetic utility of alignment gaps within intron sequences of three nuclear genes in bumble bees (*Bombus*). Mol Biol Evol. 20:87–92.

Kollman JM, Doolittle RF. 2000. Determining the relative rates of change for prokaryotic and eukaryotic proteins with anciently duplicated paralogs. J Mol Evol. 51:173–181.

Kumar S, Rzhetsky A. 1996. Evolutionary relationships of eukaryotic kingdoms. J Mol Evol. 42:183–193.

Lake JA. 2009. Evidence for an early prokaryotic endosymbiosis. Nature. 460:967–971.

Lamarck JBM. 1809. Philosophie zoologique. Paris: Dentu.

Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.

Larget B, Simon DL. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol. 16:750–759.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Lipman DJ, Altschul SF, Kececioglu JD. 1989. A tool for multiple sequence alignment. Proc Natl Acad Sci U S A. 86:4412–4415.

Lloyd DG, Calder VL. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. J Evol biol. 4:9–21.

Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol. 23:40–45.

Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol. 19:1–7.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A. 102:10557–10562.

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632–1635.

Ludwig W, Schleifer KH, Whitman WB. 2009. Revised road map to the phylum *Firmicutes*. In: De Vos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB, editors, Bergey's manual of systematic bacteriology, New York: Springer.

Lukjancenko O, Wassenaar TM, Ussery DW. 2010. Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol. 60:708–720.

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 393:162–165.

Mau B, Newton MA, Larget B. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics. 55:1–12.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48:443–453.

Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet. 30:371–403.

Nei M, Takezaki N, Sitnikova T. 1995. Assessing molecular phylogenies. Science. 267:253–254.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.

O'Malley MA, Martin W, Dupré J. 2010. The tree of life: introduction to an evolutionary debate. Biology and Philosophy. 25:441–453.

Oshima K, Kakizawa S, Nishigawa H, *et al*. (11 co-authors). 2004. Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. Nat Genet. 36:27–29.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 96:2896–2901.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol. 27:1759–1767.

Penny D, Hendy MD, Steel MA. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol Evol. 7:73–79.

Penny D, Hendy MD, Zimmer EA, Hamby RI. 1990. Trees from sequences: panacea or Pandora's box? Aust Syst Bot. 3:21–38.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Philippe H, Chenuil A, Adoutte A. 1994. Can the Cambrian explosion be inferred through molecular phylogeny? Development. 120:15–25.

Philippe H, Laurent J. 1998. How good are deep phylogenetic trees? Curr Opin Genet Dev. 8:616–623.

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

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol. 43:304–311.

Rannala B, Yang Z. 2008. Phylogenetic inference using whole genomes. Annu Rev Genomics Hum Genet. 9:217–231.

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 314:1041–1052.

Ren SX, Fu G, Jiang XG, *et al*. (39 co-authors). 2003. Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. Nature. 422:888–893.

Reyes-Prieto A, Yoon HS, Moustafa A, Yang EC, Andersen RA, Boo SM, Nakayama T, Ishida Ki, Bhattacharya D. 2010. Differential gene retention in plastids of common recent origin. Mol Biol Evol. 27:1530–1537.

Roettger M, Martin W, Dagan T. 2009. A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. Mol Biol Evol. 26:1931–1939.

Roth ACJ, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. BMC bioinformatics. 9:518.

Rzhetsky A, Nei M. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol Biol Evol. 12:131–151.

Saitou N, Nei M. 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. J Mol Evol. 24:189–204.

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

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74:5463–5467.

Sattath S, Tversky A. 1977. Additive similarity trees. Psychometrika. 42:319–345.

Schübbe S, Williams TJ, Xie G, *et al*. (11 co-authors). 2009. Complete genome sequence of the chemolithoautotrophic marine magnetotactic coccus strain MC-1. Appl Environ Microbiol. 75:4835–4852.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol. 147:195–197.

Smith TF, Waterman MS, Fitch WM. 1981. Comparative biosequence metrics. J Mol Evol. 18:38–46.

Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull. 28:1409–1438.

Stiller JW. 2011. Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. BMC Evol Biol. 11:259.

Stiller JW, Hall BD. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol Biol Evol. 16:1270–1279.

Stoye J, Evers D, Meyer F. 1998. Rose: generating sequence families. Bioinformatics. 14:157–163.

Stoye J, Moulton V, Dress AW. 1997. DCA: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. Comput Appl Biosci. 13:625–626.

Tassy P. 2010. Trees before and after Darwin. J Zool Syst Evol Res. 49:89–101.

Tateno Y, Nei M, Tajima F. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J Mol Evol. 18:387–404.

Tateno Y, Takezaki N, Nei M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol Biol Evol. 11:261–277.

Tatusov RL, Fedorova ND, Jackson JD, *et al*. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. BMC bioinformatics. 4:41.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278:631–637.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucl Acids Res. 29:22–28.

The MathWorks Inc. 2012. MATLAB. Version: 7.14. (R2012a).

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res. 22:4673–4680.

van Dongen S. 2000. Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht, The Netherlands.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 171:737–738.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

White WT, Hills SF, Gaddam R, Holland BR, Penny D. 2007. Treeness triangles: visualizing the loss of phylogenetic signal. Mol Biol Evol. 24:2029–2039.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 74:5088–5090.

Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol. 4:1286–1294.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol. 11:367–372.

Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. J Theor Biol. 8:357–366.

Danke

Mein ganz besonderer Dank gilt Prof. Dr. William Martin für die Vergabe der überaus interessanten Themen meiner Arbeit und der Möglichkeit diese Untersuchungen im Rahmen meiner Dissertation unter seiner Leitung durchzuführen. Keineswegs selbstverständlich stand er stets als Ansprechpartner für Fragen oder Diskussionen im Zusammenhang mit meiner Arbeit zur Verfügung und fand immer wieder aufmunternde Worte. Sein umfassendes Wissen, die Motivation, Begeisterung und nicht zuletzt seine Erfahrungen haben meine Arbeit begleitet und gefördert. Er ermöglichte mir außerdem zahlreiche außergewöhnliche Reisen zu internationalen Fachkonferenzen, welche meine Arbeit bereichert haben. Im besonderen möchte ich mich auch für die Einstellung als wissenschaftlicher Mitarbeiter bedanken, die mir die Aufnahme meiner Doktorarbeit erst möglich machte.

Bei Prof. Dr. Martin Lercher möchte ich mich für das Interesse an meiner Arbeit und die Bereitschaft bedanken, das Korreferat zu übernehmen.

Prof. Dr. Tal Dagan danke ich für die zahlreichen kritischen und konstruktiven Bemerkungen, ihre Ideen, den Austausch ihrer eigenen Erfahrungen als Hilfestellung und ihr mir entgegengebrachtes Vertrauen. Auch sie war jederzeit ansprechbar und stand mir mit ihrem Rat zur Seite. Ich möchte mich bei Dr. Dagan sehr herzlich für die Zusammenarbeit an allen im Rahmen dieser Arbeit entstandenen Publikationen und Tagungsbeiträgen bedanken. Sie hat diese Arbeit und insbesondere die hieraus hervorgegangenen Veröffentlichungen durch ihr Wissen, ihre Erfahrungen, ihren Enthusiasmus sowie durch ihre Entschlossenheit überaus bereichert.

I thank Dr. Giddy Landan for his patience in teaching statistical backgrounds and for his contribution to this work.

Ich bedanke mich bei allen Koautoren der im Zusammenhang mit dieser Arbeit entstandenen Veröffentlichungen.

Dr. Gabriel Gelius-Dietrich und Kathrin Hoffmann danke ich für das Korrekturlesen meiner Arbeit. Insbesondere bedanke ich mich bei Dr. Gabriel Gelius-Dietrich für seine Verbesserungsvorschläge und hilfreichen Tipps und Hinweise vor allem in Bezug auf das Softwarepaket LaTeX.

Bei meinen Bürokollegen Dr. Oliver Deusch, Dr. Xavier Pereira Brás, Kathrin Hoffmann, Christian Wöhle und Judith Ilhan bedanke ich mich außerdem für die angenehme Arbeitsatmosphäre.

Den Mitarbeitern des Instituts für Molekulare Evolution der Heinrich-Heine-Universität Düsseldorf möchte ich für die freundliche und motivierende Arbeitsatmosphäre danken.

Die vorliegende Dissertation habe ich eigenständig und ohne unerlaubte Hilfe angefertigt. Die Dissertation wurde weder in der vorgelegten noch in ähnlicher Form bei einer anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 26.04.2013

.....

(Mayo Röttger)