CHAPTER 2

Reconstructing the Universal Tree of Life

James R. Brown

Abstract

The universal tree of life depicts the evolutionary relationships of all living things by grouping them into one of three Domains of life; the Archaea (archaebacteria), Bacteria (eubacteria) and Eucarya (eukaryotes). The “canonical universal tree” topology is actually a composite of phylogenies based on single ribosomal RNA gene trees and duplicated, paralogous protein gene trees. The salient features of the canonical universal tree are: (1) all three Domains are mono/holophyletic; (2) Archaea and eukaryotes are sister groups with the Bacteria at the root; and (3) thermophilic bacteria are the earliest evolved bacterial lineage. Recent studies based on new genome sequence data suggest that the universal tree has been “uprooted” by extensive horizontal gene transfer (HGT). However, the scope of HGT is still unclear and reports of extensive trans-Domain HGT based on sequence homology, without supporting phylogenetic analysis, need careful reconsideration. Phylogenetic analysis of combined conserved proteins suggests that there is still underlying support for the concept of the universal tree.

Introduction

The universal tree of life is the depiction of the evolutionary relationships among all living organisms. The tacit supposition of the universal tree is that all living things are related genetically, however distant. Key support for this assumption comes from the subject of this book, the genetic code, which is ubiquitous with remarkably little variation. Furthermore, the basic processes of DNA replication, transcription and translation are preserved in all cells which adds support to the notion of common, if distant, origins.

While science has long attempted to classify living things, modern universal tree construction truly began with molecular evolutionary studies. Sixty years ago, Chatton and Stanier and van Niel proposed subdividing life into two fundamental groups, prokaryotes and eukaryotes (summarized in ref. 3). Later, the key features distinguishing prokaryotes from eukaryotes were better defined, namely, the lack of internal membranes (such as the nuclear membrane and endoplasmic reticulum), and replication by binary fission rather than mitosis. However, neither detailed morphology nor extensive biochemical phenotyping provided sufficient phylogenetic signal for reconstructing evolutionary relationships among prokaryotic species let alone their relationships to eukaryotes.

In the late 1970s, Woese, Fox and coworkers initiated the field of molecular prokaryotic systematics by digesting in vivo labeled 16S ribosomal RNA (rRNA) using T1 ribonuclease to produce oligonucleotide “words” then analyzing the results data using dendograms. Their rRNA dendograms showed that some unusual methanogenic “bacteria” were significant offshoots from the main bacterial clade. So deep was the split in the prokaryotes that Woese and Fox named the methanogens and their relatives “archaebacteria”, which relayed their distinctness from the true bacteria or “eubacteria” as well as met contemporary preconceptions that these...
organisms might have thrived in the environmental conditions of a younger Earth. Thus, their findings challenged the fundamental subdivision of living organisms into prokaryotes and eukaryotes thereby upsetting the assumption that evolution progressed directly from simple (prokaryotes) to more complex entities (eukaryotes).

In 1990, Woese, Kandler and Wheelis formally proposed the replacement of the bipartite prokaryote-eukaryote division with a new tripartite scheme based on three urkingdoms or Domains; the Bacteria (formally eubacteria), Archaea (formally archaebacteria) and Eucarya (eukaryotes, still the more often used name). The rationale behind this revision came from a growing body of biochemical, genomic and phylogenetic evidence which, when viewed collectively, suggested that the Archaea were unique from eukaryotes and the Bacteria. The discovery of the Archaea was a significant event, which added a new dimension to the construction of the universal tree since evolutionary relationships between the three major subdivisions had to be considered (Fig. 1).

### Topology of the Universal Tree

The obvious challenge in universal tree reconstruction is determining which Domain evolved first and, therefore, is the root of the universal tree. Assuming that each Domain is monophyletic there are three possible answers (depicted respectively in Fig. 2) (1) Bacteria diverged first from a lineage producing Archaea and eukaryotes (AE tree) or (2) eukaryotes diverged from a fully prokaryotic clade, consisting of Archaea and Bacteria (AB tree) or (3) the Archaea diverged first such that Bacteria and eukaryotes (BE tree) are sister groups.

In terms of species diversity and carbon biomass, the Archaea are far from insignificant. Early interest in the Archaea was motivated by their remarkable success in flourishing in the harshest of environments, which earned them the title of "extremophiles". However, more
Figure 2. Three possibilities for the rooting of the universal tree. A) Bacteria diverged first from a lineage producing Archaea and eukaryotes (called here the AE tree); B) Eukaryotes diverged from a fully prokaryotic clade, consisting of Bacteria and Archaea (the AB tree) or; C) the Archaea diverged first such that eukaryotes and Bacteria are sister groups (the BE tree).

Recent studies show that many archaeal species are "mesophiles", living in oceans, lakes, soil, and even animal guts.9

Prior to whole genome sequence data, considerable knowledge had accumulated on the comparative biochemistry, and cellular and molecular biology of the Archaea (for a review see refs. 10-13). Archaea seem to have a few unique biochemical and genetic traits as well as a variety of metabolic regimes, which deviate from known metabolic pathways of Bacteria and eukaryotes, and are not simply particular environmental adaptations. Recent genome comparisons found 351 archaea-specific "phylogenetic footprints" or combinations of genes uniquely shared by two or more archaeal species but not found in either bacteria or eukaryotes.14 However, such inventories might over estimate the number of unique functional proteins since hyperthermophilic Archaea and Bacteria tend to have more split genes compared to their mesophilic counterparts.15 Archaeal and bacterial species are definitely prokaryotes with generally similar ranges of cell sizes, genes linked in operons, large circular chromosomes often accompanied by one or more smaller circular DNA plasmids, and lacking nuclear membranes and organelles.

However, Archaea and eukaryotes share significant components of DNA replication, transcription, and translation, which are either not found in Bacteria or replaced by an evolution-
The Genetic Code and the Origin of Life

Cenancestor

Gene A

Gene Duplication

Gene A1

Gene A2

Eukaryotes

Archaea

Bacteria

Figure 3. Conceptual rooting of the universal tree using paralogous genes. Gene A was duplicated in the cenancestor such that all extant organisms have paralogous copies, gene A1 and gene A2. The two genes are sufficiently similar to allow for the construction of reciprocally rooted trees thus rooting the tree of one paralog with that of the other. The topology depicted here, Archaea and eukaryotes as sister groups with the root in Bacteria, has been consistently supported by paralogous trees (see text).

ary unrelated (analogous) enzyme. Many DNA replication and repair proteins are homologous between Archaea and eukaryotes but completely absent in Bacteria. While the archaeabacterium, *Pyrococcus abyssi*, was recently shown to have a bacteria-like origin of DNA replication, most of its replication enzymes are eukaryote-like. Archaeal DNA scaffolding proteins are remarkably similar to eukaryotic histones. Eukaryotes and the Archaea have similar transcriptional proteins, such as multi-subunit DNA-dependent RNA polymerases, as well as sharing translation initiation factors not found in the Bacteria. Thus, based on cellular and genetic components, the Archaea seem to occupy a middle ground between the Bacteria and eukaryotes, a conclusion which serves little in resolving the rooting problem. Only in molecular phylogenetics lies such hope.

The lack of an outgroup to all living things meant that the rooting of the universal tree could only be resolved by using paralogous genes to construct reciprocally rooted trees (Fig. 3). Iwabe and coworkers aligned amino acids from five conserved regions shared by the elongation factors (EF) Tu/1α and EF-G/2 genes of the archaeabacterium, *Methanococcus vannielii*, and several species of Bacteria and eukaryotes. According to protein sequence similarity and neighbor-joining trees, both EF-1α and EF-2 genes of Archaea were more similar to their respective eukaryotic, rather than bacterial, homologs. Gogarten and coworkers developed composite trees based on duplicated ATPase genes where the V-type A and V-type B occurs in Archaea and eukaryotes and the F0F1-type β and F0F1-type α occurs in Bacteria. In agreement with the elongation factor rooting, reciprocally rooted ATPase subunits trees also showed that the Archaea, represented by a sole species *Sulfolobus acidocaldarius*, were closer to eukaryotes than to Bacteria.

Subsequent paralogous protein rootings based on aminoacyl-tRNA synthetases and carbamoylphosphate synthetase confirmed the rooting in the Bacteria and linking Archaea
and eukaryotes as sister groups. If one argues that enzymes involved in DNA replication, transcription and translation, so-called "information" genes, are core to living things then the evolutionary scenario suggested by paralogous gene trees seems particularly reasonable. Thus emerged the "canonical" universal tree with the Archaea and eukaryotes being sister groups, the rooting in the Bacteria, and all three Domains as monophyletic groups.

**Uprooting the Universal Tree**

Despite the convincing results from paralogous gene trees, the rooting of the universal tree has not been without controversy. Phylogenetic analyses using alternative methods and expanded data sets raised questions about the rooting of the universal tree and the monophyly of the Archaea. Philippe and coworkers have maintained that phylogenies of distantly related species are strongly affected by saturation for multiple mutations at nearly every amino acid position in a protein. Unequal mutation rates between different species can lead to long branch attraction effects. However, a greater issue is the degree to which horizontal gene transfers between the Domains of life have affected the actual viability of constructing a definitive universal tree.

The increasing size of sequence databases adds to the species richness of universal trees. Perhaps not surprisingly, nature provides plenty of exceptions to the canonical universal tree paradigm. In most cases, the key hypothesis invoked has been horizontal gene transfer or HGT. Simply stated, HGT is the exchange of genes between organisms which are not directly related by evolutionary descent. Many examples of HGT between closely related species are known, such as the transfer of bacterial antibiotic resistance genes. The extent and nature of more ancient HGT events, (i.e., trans-Domain HGT between species of one Domain to species of another Domain), is an important and open evolutionary question which is further considered for the remainder of this chapter.

Among the first documented trans-Domain HGT events involved ATPase subunits which were actually key in rooting the universal tree. Archaeal V-type ATPases were reported for two bacterial species, *Thermus thermophilus* and *Enterococcus hirae*, while a bacterial F1-ATPase β subunit gene was found in the Archaea, *Methanosarcina barkeri*. Consequently, Forterre and coworkers suggested that the ATPase subunit gene family had not been fully determined, and that other paralogous family members might be discovered Hilario and Gogarten believed that the observed distribution of ATPase subunits was the result of a few, rare HGTs. In support of the latter view, broader surveys have failed to detect archaeal V-type ATPases in other bacterial species.

The HGT debate was amplified by a growing number of examples where single gene trees, although not uniquely rooted, had irreconcilable topologies to that of the canonical universal tree. In 1995 Golding and Gupta examined the phylogenetic trees for 24 universally conserved proteins and found only nine with the AE tree topology. Although subsequent phylogenetic analyses by Gupta and Golding and Roger and Brown slightly modified the number of protein trees with AE topologies, a significant number of proteins still conflicted with the canonical universal tree. Feng, Cho and R.F. Doolittle found that in the 34 universal protein trees they constructed, AE, AB, and BE topologies occurred for 8, 11, and 15 proteins, respectively. A broader survey involving phylogenetic analysis of 66 proteins found that AE, AB, and BE topologies occurred for 34, 21, and 11 protein trees, respectively, with the remaining trees having indeterminate relationships among the Domains. New genome sequence data have further reduced the AE list with additional examples of horizontal gene transfer between eukaryotes and bacteria, such as isoleucyl-tRNA synthetases.

**Genomes and HGT**

Genomes are being sequenced at a remarkable pace, the progress of which can be followed at number of websites including those of the NCBI Genome (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/bact.html) and TIGR Microbial (http://www.tigr.org/tdb/mdb/mdb.html) Databases. This new abundance of sequence data has resulted in a more, not less, confusing
picture of the universal tree. Comparative analysis of archaea, bacterial and eukaryotic genomes suggest that relatively few genes are entirely conserved across all genomes. Important biochemical pathways appear to be incomplete in some organisms. In some instances, a protein has been discovered to take over the catalytic role of an unrelated protein, so-called nonorthologous gene replacement.\(^4\)

Phylogenetic analyses of conserved proteins suggest that trans-Domain HGT has been extensive. Lake and colleagues suggest that based on their propensity for HGT, genes could be divided into two categories, informational and operational genes.\(^5\) Informational genes, which include the central components of DNA replication, transcription and translation, are less likely to be transferred between genomes than operational genes involved with cell metabolism. The fact that informational gene products, at least qualitatively, have more complex interactions might restrict their opportunities for genetic exchange and fixation.\(^6\) Additional support for this view is the conservation of genomic context for translation-associated genes in bacteria.\(^7\)

Despite their critical role in protein synthesis and ancient origins (without them interpretation of the genetic code would be impossible), aminoacyl-tRNA synthetases have been extensively shuttled between genomes (for a review see refs. 53-55). Phylogenetic trees suggest that class I isoleucyl-tRNA synthetases may have been transferred from an early eukaryote to bacteria as a specific adaptation to resist a natural antibiotic compound.\(^8\) Orthologous genes to eukaryotic glutaminyl-tRNA synthetase occur in many proteobacteria and \emph{D. radioutrarums} but not in other Bacteria or the Archaea.\(^9\) Archaea and some bacteria, Spirochaetes, share novel type of lysyl-tRNA synthetases\(^7\) and phenylalanyl-tRNA synthetases.\(^9\)

Metabolic genes can have surprising species distributions such as the mevalonate pathway for isoprenoid biosynthesis. The mevalonate pathway has been well studied in humans because 3-hydroxy-3-methylglutaryl coenzyme A [HMG-CoA] reductase is the target for the statin class of cholesterol-lowering drugs. The mevalonate pathway was long believed to be specific to eukaryotes since most bacteria utilize an evolutionary unrelated metabolic route for isoprenoid biosynthesis, the pyruvate/GAP pathway. However, recent genome surveys and phylogenetic analyses have found not only HMGCoA reductase but also four other enzymes in the mevalonate pathway in Gram-positive coccal bacteria.\(^10\)-\(^12\) The genes are also found in the Archaea and the bacterial spirochaete, \emph{Borrelia burgdorferi}. However, the mevalonate pathway is absent from the completely sequenced genome of a closely related Spirochaete, \emph{Treponema pallidum}, and the Archaea have likely substituted an analogous protein for at least one enzyme in the pathway.\(^13\) In those Bacteria with the mevalonate pathway, the genes encoding component enzymes are tightly linked suggesting that all genes might have been transferred simultaneously. Genes contributing products to a common metabolic pathways might be more readily fixed in the recipient genome than isolated, individual genes, which, in turn, would favor the organization of pathway genes into tightly linked operons.\(^14,15\)

\section*{Cautionary Notes on the HGT Hypothesis}

Recent science news reports have painted the picture that significant fractions of the scientific community engaged in genomics and universal tree studies have taken "a sky is falling" attitude towards the possibility of reconstructing cellular evolution in light of widespread HGT.\(^16,17\) In summary, their view is that while phylogenetic approaches are still useful for mapping the evolution of individual proteins, HGT has significantly confounded the reconstruction of the universal tree, hence, any discerned patterns in early genome evolution are suspect.\(^18\) However, there is a need to critically evaluate methods for detecting HGT, which in some cases, can lead to overestimates of its occurrence.\(^19,20\)

Reports of HGT without supporting phylogenetic analyses should be carefully scrutinized. Comparative studies based on BLAST\(^21\) analyses have concluded that HGT has extensively occurred between Archaea and Bacteria. Koonin and coworkers\(^22\) found that 44% of the gene products of the archaeabacterium, \emph{Methanococcus jannaschii} were more similar to bacterial over...
Reconstructing the Universal Tree of Life

Figure 4. Detection of horizontal gene transfer (HGT) from phylogeny. Hypothetical protein trees for three bacterial species (B1-B3) and three archaeal species (A1-A3). A) The true rooting of the tree postulates a split between the Archaea and Bacteria, which results in two monophyletic clusters. B) The lowest branching bacterial species, B1, has a more rapid rate of amino acid substitution than other bacterial species which results in phylogenetic programs as well as homology searching software implicating the Archaea as the closest relatives. At first glance, the tree would suggest HGT between B1 and Archaea. However, the clustering of species is actually the result of the new position of the root, which was shifted by the "attraction" of the B1 branch to the outgroup, the Archaea. C) Strong phylogenetic evidence for HGT is the "imbedding" of a distantly related in-group species within the outgroup and away from the root. In this example, bacterial species B1 clusters with a more derived archaeal species, A3, which strongly suggests HGT occurred from the Archaea (A3) to Bacteria (B1).

eukaryotic proteins while only 13% were more like eukaryotic proteins. Nelson and coworkers\(^7^2\) reported that 24% of proteins from *Thermotoga maritima*, a thermophilic bacterium with a deep rRNA tree lineage, were most similar to archaeal proteins. However, deep branching species of one Domain are susceptible to arbitrary clustering with species from the other Domains, such as bacterial thermophiles with the Archaea and eukaryotes.\(^7^2\) Differences in evolutionary rates can lead to an incorrect rooting which will result in mistaken occurrences of HGT between the deep branching species and the outgroup (Fig. 4A and 4B). Conversely, protein trees where an in-group species is solidly embedded within an outgroup clade provide strong evidence for HGT (Fig. 4C). Consequently, phylogenetic analysis suggests that *T. maritima* received far fewer genes from the Archaea than first estimated by homology searches.\(^7^2\)\(^7^3\) Phylogenetic analyses of putative archaeal-like proteins from *Deinococcus radiodurans*, a bacterium which branches nearly as deeply as *Thermotoga* in rRNA trees, suggests that HGT involving either Archaea or eukaryotes occurred for fewer than 1% of its total genome complement.\(^7^4\)

Some remarkable claims of direct HGTs from bacteria to vertebrates were made in the historic publication of the first draft of the human genome sequence by the International Human Genome Sequencing Consortium (IHGSC) in 2001.\(^7^5\) In the paper, they stated that as many as 113 vertebrate genes, some only found in humans, were the result of direct HGT from bacteria. This conclusion was based on BLASTP score analyses where the expect value (E-values) of human gene matching a bacterial gene was 9 orders of magnitude greater than the value to the closest related nonvertebrate eukaryote gene. The possibility of direct bacteria to vertebrate HGT has several important evolutionary and medical ramifications. First, any gene transferred and fixed in the genome of a multicellular organism, like vertebrates, would need to be introduced into the germ cell line. Second, bacterial genes could only be functionally expressed in vertebrate genomes if they could readily adapt to the eukaryotic gene regulon. Finally, there are serious public health concerns if the human gene pool could become permanently contaminated from bacterial genes as a consequence of infection or the ingestion of genetically modified foods. However, three independent studies concluded that there was no evidence for HGT from bacteria to vertebrates.\(^7^6\)\(^7^8\)
In our study, we examined all 28 cases where the IHGSC had verified the presence of the gene in the human genome by PCR, BLAST searches of additional databases, in particular nonvertebrate EST databases (i.e., the National Center for Biotechnology Information “EST others” database), revealed many homologs in nonvertebrates (i.e., fungi, nematodes and insects) which were previously undetected. In other instances, a nonvertebrate homolog was found in public databases but at a threshold above the E-value cut-off of 9 orders of magnitude used in the IHGSC study. However, alignment of multiple sequences followed by phylogenetic analyses, resulted in monophyletic clades of eukaryotes with both vertebrates and nonvertebrates together. Of the 28 genes examined, only one instance of possible vertebrate to bacteria HGT was found. There was no evidence of bacteria to vertebrate HGT.

Hypothetical HGT events have also been suggested by analysis of differences in nucleotide composition (G+C content) between donor and recipient coding regions. However, intragenomic base composition can be highly variable between chromosomal regions which could lead to over estimates in the number of transferred genes. Arguably, genes might be more likely to be transferred in clusters, such as operons, particularly if the genes encode several proteins in a common biochemical pathway. Thus, patterns of gene position or context across genomes might be useful indicators of HGT. However, even simple operons can vary greatly among closely related species or be identical among highly unrelated ones. An example is the organization of the two genes coding the alpha and beta subunits of phenylalanyl-tRNA synthetase which are cotranscribed in most species of Bacteria and Archaea but have become dispersed in the genomes of others through what appears to be multiple, independent events.

In summary, reports of HGT need to be critically evaluated. Proper scientific inquiry should begin with the assumption of the null hypothesis, which, in the case of comparative genomic studies, is that HGT has not occurred and that all genes evolved by direct inheritance. Only after adopting such a stance, can we begin to grasp the true role of HGT in genome evolution.

**Possible HGT Patterns and Processes**

In addition to the detection of Domain HGT, there are issues about the magnitude, directionality and timing of this phenomena are discussed below in the context of the three possible topologies of the universal tree.

First, trees which depict Archaea and eukaryotes as sister groups (the AE tree in Fig. 2) largely result from the phylogenetic analyses of proteins involved in DNA replication, transcription and translation. Archaea seem to utilize a wider range of eukaryote-type proteins for these processes than Bacteria. Paralogous gene trees also position Archaea and eukaryotes as sister groups although it has been suggested that such results are idiosyncratic due to more rapid rates of evolutionary change in Bacteria.

Among the three possible universal tree scenarios, only trees with the AE clustering depict, even if occasionally, all three Domains to be monophyletic simultaneously. If extensive polyphyly (species from different Domains in the same clade) is evidence for HGT then, by default, monophyly indicates evolution in the absence of HGT. Given the large universe of genes, Domain monophyly appears to be a rare occurrence. However, the existence of some monophyletic gene trees should suggest that their topology reflects the underlying evolutionary trajectory of the species involved without the complication of HGT. If true, then the overall scenario of cellular evolution, heavily diluted by HGT events, remains the canonical universal tree with a rooting in the Bacteria with Archaea and eukaryotes as sister groups. However, the persistence of monophyly in universal trees is highly dependent upon the diversity of species sampled. Notably, genome sequences from simple, single-cell eukaryotes will likely reveal instances of Domain HGT previously unnoticed in higher eukaryotes.

Second, there are phylogenies where Archaea and Bacteria are closest relatives (the AB tree in Fig. 2). However, in those trees, one or both Domains are always para/polyphyletic groups. Such tree topologies are evidence for HGT between Archaea and Bacteria, the patterns for which can be often complex. The genes and species implicated in Archaea-Bacteria HGT are
highly varied. Glutamine synthetases, glutamate dehydrogenase and HSP70 of Archaea are closely related to orthologs from Gram-positive bacteria. Hyperthermophilic archaeal and bacterial species share a reverse gyrase which is likely a common adaptation to life at extremely high temperatures. Catalase-peroxidase genes appear to have been exchanged between Archaea and pathogenic proteobacteria. Two component signal transduction systems in the Archaea as well as fungi and slime molds were likely acquired from the Bacteria. However, as discussed above, similarities between Bacteria and Archaea are not always conclusive evidence for HGT events. Species forming low branches in the two Domains can be attracted or cluster together because of rooting artifacts. In addition, gene distributions shared by Bacteria and Archaea but not eukaryotes might be caused by gene loss or replacement in eukaryotes rather than HGT between Archaea and Bacteria.

The third universal tree topology, Bacteria and eukaryotes as closest relatives or the BE tree (Fig. 2), might result from specific bi-directional gene transfers. Some bacterial species appear to have acquired genes from eukaryotes such as the glutaminyl-tRNA synthetase gene. On the other hand, eukaryotes have likely integrated a large number of bacterial genes as a consequence of endosymbiosis related to mitochondria and plastid biogenesis. The endosymbiosis theory of organelle origins is a widely accepted fact. However, the deeper consequences of endosymbiosis to eukaryotic genome evolution are just being revealed by genome sequencing projects. Genome comparisons and phylogenetic analyses involving Arabidopsis thaliana and Synechocystis sp., suggest that plants obtained from 1.6% (-400 genes) to 9.2% (-2200 genes) of their gene complement from cyanobacteria, the bacterial progenitor of plastids. Phylogenies for many conserved proteins, such as the glycolytic pathway enzymes suggest bacterial origins for many eukaryotic genes (for a review see ref. 13). The occurrence of mitochondria-targeted genes in simple protists which both lack mitochondria (amitochondrial) and appear as early evolved eukaryotic lineages, suggests endosymbiotic transfer of genes to the nuclear genome occurred early in the evolution of eukaryotes. In some instances, the organelle gene has either contributed a new function or replaced the original orthologous gene in the genome of the host. However, other phylogenetic trees, namely of aminoacyl-tRNA synthetases, suggest that patterns of integration of bacterial genes in the eukaryotic genome via endosymbiosis might be more complex.

**Universal Trees Based on Multiple Datasets**

Construction of universal trees based on the distribution of genes is a logical use of genomic sequence data in evolutionary biology. The underlying principal of this approach is that species with the largest proportion of common genes should be more recently diverged than species with fewer shared genes. There are several important methodological considerations such as distinguishing orthologous genes from paralogous ones, accurate prediction of genes, and normalization of gene inventories across genomes. Although employing somewhat different approaches, studies which constructed universal trees from gene distributions generally found tree topologies remarkably similar to that of the canonical universal tree and rRNA tree. However, it has been argued that while genome inventories might tell us about the similarities in the contents of genomes from different species, the nuisances of HGT involving universally conserved genes are lost.

Potentially, gene order could also be used to reconstruct phylogenies of bacteria and archaea since many recognizable operon organizations occur across these two Domains. However, gene order is poorly conserved between species and is unlikely to be a useful phylogenetic marker although overall neighborhoods of genes on the chromosome might be preserved because of functional and regulatory consequences.

On the other hand, the combination or concatenation of multiple protein datasets derived from genome sequences might be useful for the phylogenetic reconstruction of universal trees. Phylogenies based on concatenated protein datasets are potentially more robust and representative of the evolutionary relationships among species since the number of phylogenetically
informative sites and sampled gene loci are greatly increased. The main principle behind combining data is that it allows for the amplification of phylogenetic signal, and increased resolving power, in cases where signal is masked by homoplasy (similarities in amino acids for reasons other than inheritance) among the individual gene data sets. Such protein datasets have helped resolve evolutionary relationships among photosynthetic bacteria and eukaryotic protists.

By definition, a universally conserved protein occurs in every organism. The increasing number of completely sequenced genomes will invariably lead to the shrinking of this inventory since the odds will increase for finding exceptional cases. For example, the 70 kilo-Dalton heat shock protein (HSP70), once thought to be highly conserved from the perspective of both amino acid substitutions and species distribution, is absent from several species of Archaea. In many cases, the biochemical function is still required but an evolutionary unrelated enzyme serves as the catalyst. Arguably, only those proteins found in all completely sequenced genomes are conserved enough to provide a continuous picture of all lineages back to the last universal common ancestor. Fortunately, the contemporary collection of completely sequence genomes represents fairly diverse groups of Bacteria, Archaea and eukaryotes. Therefore, for purposes of universal tree reconstruction, the list of completely conserved proteins across the three Domains is unlikely to be further reduced with new genomes.

Recently, we constructed universal trees based on the combined alignments of proteins conserved across 45 species from all three Domains. Proteins were selected on fairly strict criteria of being conserved across all species and being orthologous (i.e., paralogs or duplicated proteins within a species were eliminated from the entire analysis). For eukaryotes, where two copies of a gene might exist, one targeted to the mitochondria and the other to the cytoplasm, only the latter was used since the cytoplasmic version best tracks the evolution of the eukaryotic nucleus. The determined number of conserved proteins, 23, was far fewer than previous genomic studies. For example, the Clusters of Orthologous Groups of proteins (COGs) database reports for 34 complete genomes, a total of 78 completely conserved proteins. However, we included several additional genomes, a few which were incomplete at the time of the study. In addition, if the collection of organisms is diverse, then the likelihood increases that particular lineages, by chance, have lost a particular pathway or replaced components with analogous proteins. Our list, shown in Table 1, represents the most highly conserved or widely found proteins known to date. The edited multiple sequence alignment of the concatenated dataset of 23 proteins was 6591 amino acids in length, which was far larger than any single protein dataset, and is the largest applied to universal tree reconstruction.

Similar to universal rRNA trees, all combined protein dataset phylogenetic trees strongly supported the monophyly of the three Domains. On average, archaeal and eukaryotic species were slightly more similar to each other than either was to Bacteria. However, it cannot be confirmed that Archaea and Eucarya share a last common ancestor since the tree is unrooted. Within each Domain, the branching order of most nodes are well supported by bootstrap replications (> 70%). Although fewer genomes of Archaea and eukaryotes have been completely sequenced, branching orders of those species were consistent with contemporary views of organism evolution.

In the Bacteria, the major subdivisions of Bacillus/Clostridium (low G+C Gram positives), Spirochaetes, and Proteobacteria were strongly supported as being monophyletic, as postulated by the universal rRNA trees. However, a major departure was the placement of Spirochaetes (represented by the species Treponema pallidum and Borrelia burgdorferi) as the first bacterial branch rather than thermophiles (Aquifex aeolicus and Thermotoga maritima). While the basal position of Spirochaetes is incompatible with hypotheses regarding the thermophilic origins of life, there are suggested instances of HGT between Spirochaetes and Archaea, such as class I lysyl-tRNA synthetases. In the combined protein alignment phylogenetic method, the inclusion of such proteins would tend to move the Spirochaete branch to a more basal position in the bacterial clade.
Table 1. Proteins included in concatenated alignments, the number of residues, and the support for domain monophyly in individual protein trees

<table>
<thead>
<tr>
<th>Cellular Function</th>
<th>Protein Name</th>
<th>Number of Amino Acids</th>
<th>Support for Domain Monophyly</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Archaea</td>
<td>Bacteria</td>
</tr>
<tr>
<td>1 translation</td>
<td>alanyl-tRNA synthetase</td>
<td>502</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>aspartyl-tRNA synthetase&lt;sup&gt;c&lt;/sup&gt;</td>
<td>249</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>glutamyl-tRNA synthetase&lt;sup&gt;c&lt;/sup&gt;</td>
<td>188</td>
<td>50 (-)</td>
</tr>
<tr>
<td>4</td>
<td>histidyl-tRNA synthetase</td>
<td>166</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>isoleucyl-tRNA synthetase</td>
<td>552</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>leucyl-tRNA synthetase&lt;sup&gt;c&lt;/sup&gt;</td>
<td>358</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>methionyl-tRNA synthetase</td>
<td>306</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>phenylalanyl-tRNA synthetase b subunit</td>
<td>177</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>threonyl-tRNA synthetase</td>
<td>305</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>valyl-tRNA synthetase</td>
<td>538</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>initiation factor 2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>337</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>elongation factor G&lt;sup&gt;c&lt;/sup&gt;</td>
<td>536</td>
<td>64(87)</td>
</tr>
<tr>
<td>13</td>
<td>elongation factor Tu&lt;sup&gt;c&lt;/sup&gt;</td>
<td>340</td>
<td>- (42)</td>
</tr>
<tr>
<td>14</td>
<td>ribosomal protein L2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>192</td>
<td>46(-)</td>
</tr>
<tr>
<td>15</td>
<td>ribosomal protein S5&lt;sup&gt;c&lt;/sup&gt;</td>
<td>131</td>
<td>46(19)</td>
</tr>
<tr>
<td>16</td>
<td>ribosomal protein S8&lt;sup&gt;c&lt;/sup&gt;</td>
<td>118</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>ribosomal protein S11&lt;sup&gt;c&lt;/sup&gt;</td>
<td>110</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>aminopeptidase P</td>
<td>95</td>
<td>-</td>
</tr>
<tr>
<td>19 transcription</td>
<td>DNA-directed RNA polymerase b chain&lt;sup&gt;c&lt;/sup&gt;</td>
<td>537</td>
<td>99(78)</td>
</tr>
<tr>
<td>20 DNA replication</td>
<td>DNA topoisomerase I&lt;sup&gt;c&lt;/sup&gt;</td>
<td>236</td>
<td>-</td>
</tr>
<tr>
<td>21 metabolism</td>
<td>DNA polymerase III subunit&lt;sup&gt;c&lt;/sup&gt;</td>
<td>194</td>
<td>46(49)</td>
</tr>
<tr>
<td>22</td>
<td>signal recognition particle protein&lt;sup&gt;c&lt;/sup&gt;</td>
<td>298</td>
<td>71(39)</td>
</tr>
<tr>
<td>23</td>
<td>rRNA dimethylase full alignment length&lt;sup&gt;d&lt;/sup&gt;</td>
<td>126</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>truncated alignment length&lt;sup&gt;e&lt;/sup&gt;</td>
<td>6591</td>
<td>-</td>
</tr>
</tbody>
</table>

<sup>a</sup> Length of alignments after removing ambiguously aligned regions. <sup>b</sup> Occurrence of monophyletic nodes in 100 bootstrap replicated datasets of protein distance/neighbor-joining and maximum parsimony methods (in parentheses where maximum parsimony values differ from those of the neighbor-joining consense tree). Dash indicates that the nodes were not monophyletic. <sup>c</sup> Proteins included in both the full and truncated alignments. <sup>d</sup> Length of multiple sequence alignment, which included all proteins, used to produce phylogeny in Figure 5. <sup>e</sup> Length of multiple sequence alignment, which excluded proteins where the Bacteria were not monophyletic, used to produce phylogeny in Figure 6. Table adapted from ref. 107.

Examination of the individual gene trees revealed topologies where the Domains, primarily the Bacteria, were not monophyletic thus implicating possible instances of HGT (Table 1). Interestingly, none of the 23 individual protein trees suggested that hyperthermophilic bacteria, the species *Thermotoga maritima* and *Aquifex aeolicus*, exchanged genes with either eukary-
Figure 5. Universal tree based on 23 combined protein datasets. Minimal length maximum parsimony universal tree based on 23 combined protein datasets is shown. Spirochaetes are placed as the lowest branching Bacteria. Numbers along the branches show the percent occurrence of nodes in 50% or greater of 1000 bootstrap replicates of maximum parsimony (plain text) and neighbor joining (italicized text) analyses or 1000 quartet puzzling steps of maximum likelihood analysis (in parentheses). Dashed lines show occasional differences in branching orders in neighbor-joining trees. Scale bar represents 100 amino acid residue substitutions. CFB stands for the Cytophaga-Flexibacter-Bacteroides group of bacteria. For a full explanation of methods of construction see ref. 107. Figure adapted from ref. 107.

In contrast to the combined alignment of 23 proteins, phylogenetic trees based on the alignment of 14 non-HGT proteins agreed with universal rRNA trees in the placement of hyperthermophilic species, A. aeolicus and T. maritima, as the lowest branching bacterial lineages while Spirochaetes were a derived group (Fig. 6). However, high G+C and low G+C Gram-positives were not collectively monophyletic as previously reported for rRNA and other molecular markers. The clustering of Chlamydiales, CFB and Spirochaetes together is also novel relative to rRNA trees. The agreement between the dataset that excluded horizontal
transferred genes (truncated protein tree) and the rRNA tree, in the placement of extreme thermophiles as the basal lineage in the Bacteria lends further support to the theory that life evolved at high temperatures.\textsuperscript{110-112} However, there are still many unresolved issues surrounding the "hot" origin of life hypothesis such as the maintenance of extracellular biochemical reactions\textsuperscript{18} and the stability of RNA molecules at extreme temperatures.\textsuperscript{113}

Genes found only in thermophilic Bacteria and Archaea are just as likely to be shared syplexomorphies, which were later lost in other bacterial species. Truncated protein trees showed a fundamental division in the Bacteria where, after diverging from hyperthermophiles, Proteobacteria split from all other bacteria. Furthermore, within the Proteobacteria, the earliest diverged group is the alpha-subdivision, represented by \textit{Rickettsia prowazekii}, from which the endosymbiont progenitor of the mitochondria likely evolved.\textsuperscript{114,115} The early emergence of alpha-Proteobacteria suggests that endosymbiotic relationships between eukaryotes and bacteria could have occurred early in cellular evolution, perhaps shortly after the divergence of the Domains Bacteria, Archaea and eukaryotes. As bacterial species were evolving, they could have shared genes with early eukaryotes either directly or through secondary transfers with free-living relatives of endosymbionts. The net result would be the seemingly extensive exchange of genes between eukaryotes and many diverse, now distantly related, groups of bacteria.

Phylogenetic analysis of combined protein datasets perhaps represents an important approach in the utilization of genome sequence data to address evolutionary questions. While HGT has likely played an important, if not fully defined, role in cellular evolution perhaps genomes have retained sufficient phylogenetic signal for the reconstruction of meaningful universal trees.

In addition, phylogenetic analysis of combined protein and/or nucleotide alignments might be a useful alternative to phylogenetic analysis of rRNA molecules in bacterial systematics. While some analyses suggest the phylogenetic signal for combinations of certain conserved proteins within the Bacteria might be low,\textsuperscript{55,116} other studies based on wider collections of proteins support new relationships among bacterial groups.\textsuperscript{102}

\textbf{Concluding Remarks}

The apparent occurrence of extensive HGT across the Domains of life has prompted much speculation on its significance to early cellular evolution. Networks of genetic interactions at the base of the universal tree have been suggested to be so intense as to render useless the concept of a single cellular ancestor for contemporary lineages.\textsuperscript{41,117} Other radical positions discuss the emergence of eukaryotes from the complete fusion of genomes from an archaeabacterium and bacterium (for a review see ref. 13). Martin and Müller\textsuperscript{118} proposed a more stepwise progression to eukaryotes beginning with a hydrogen-dependent host, likely an archaeabacterium, and a respiring bacterial symbiont. W.F. Doolittle\textsuperscript{119} suggests a ratchet-like addition of bacterial content to the eukaryotic genomes from either a prokaryotic food source or gene transfers as a consequence of multiple but brief endosymbiotic associations. Such controversies will either be resolved or amplified as genomes from more taxa are sequenced. While HGT has certainly unsettled the universal tree of life, it is premature to say that the tree has been permanently uprooted.\textsuperscript{121}

\textbf{Acknowledgments}

Preliminary sequence data reported in ref. 107 and used to construct the trees in Figures 5 and 6 were obtained from various public databases. \textit{Chlorobium tepidum}, \textit{Enterococcus faecalis}, \textit{Porphyromonas gingivalis}, and \textit{Streptococcus pneumoniae} sequence data were obtained from The Institute for Genomic Research (TIGR) through the website at http://www.tigr.org and were funded by the U.S. Department of Energy (DOE), the National Institute of Allergy and Infectious Disease (NIAID) of NIH, TIGR, and the Merck Genome Research Institute. Preliminary sequence data for \textit{Actinobacillus actinomyetemcomitans} were obtained from the \textit{Actinobacillus} Genome Sequencing Project, University of Oklahoma ACGT and B.A. Roe, F. Z. Najar, S. Clifton, Tom Ducey, Lisa Lewis and D.W. Dyer through the website http://
www.genome.ou.edu/act.html which was supported by USPHS/NIH grant from the National Institute of Dental Research. Preliminary sequence data for *Pyrococcus furiosus* were obtained from the Utah Genome Center, Dept. of Human Genetics, University of Utah through the website http://www.genome.utah.edu/sequence.html which was supported by DOE. Preliminary sequence data for *Streptomyces coelicolor* were obtained from The Sanger Center through the website at http://www.sanger.ac.uk/Projects/S_coelicolor/.

References

32 The Genetic Code and the Origin of Life

The Genetic Code and the Origin of Life
Ribas de Pouplana, L. (Ed.)
2004, XVI, 254 p., Hardcover