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[1] and Stanier and van Niel[2] 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 recticulum), and replication by binary fission rather than mitosis.[4,5] 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.[6] So deep was the split in the prokaryotes that Woese and Fox[7] 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

The Genetic Code and the Origin of Life, edited by Lluís Ribas de Pouplana.

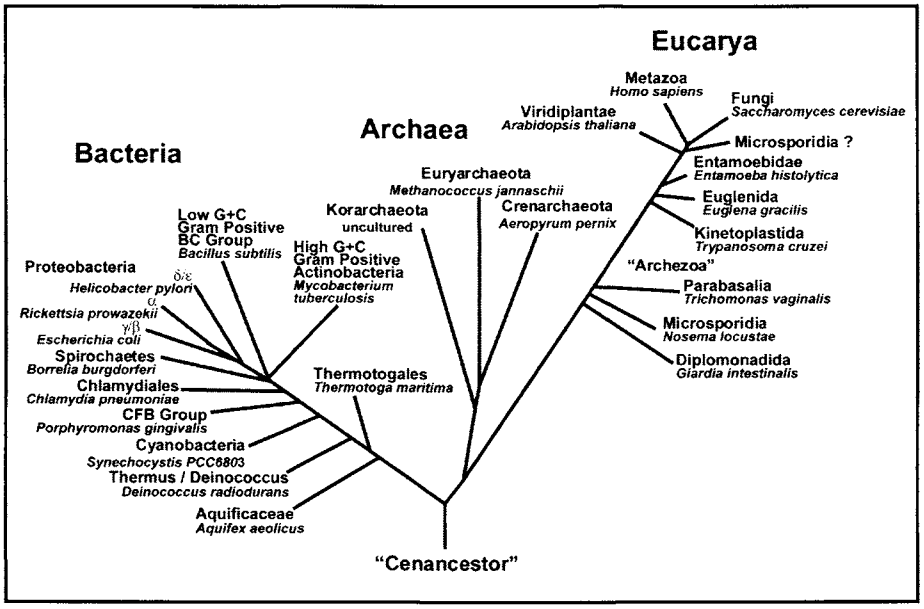

Figure 1. Schematic drawing of the universal tree showing the relative positions of evolutionary pivotal groups in the domains Bacteria, Archaea, and Eucarya. The phylum or other higher order name is given for key groups of organisms with a representative species named in italics below. The location of the root (the cenancestor) corresponds with that proposed by reciprocally rooted gene phylogenies (see text). The question mark beside Microsporidia denotes recent suggestions that it might branch higher in the eukaryotic portion of the tree.[120] (Branch lengths have no meaning in this tree). Figure adapted from ref. 13.

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[8] 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

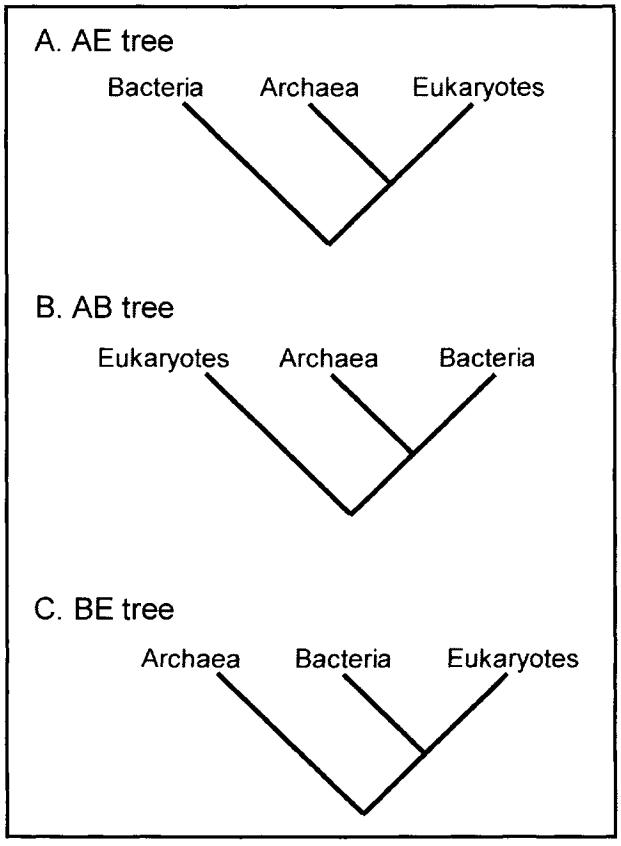

Figure 2. Three possibilities for the rooting of the universal tree. A) Bacteria diverged first from a lineage producing Archaea and eukaryotes (called here the AE tree); B) Eukaryotes diverged from a fully prokaryotic clade, consisting of Bacteria and Archaea (the AB tree) or; C) the Archaea diverged 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-

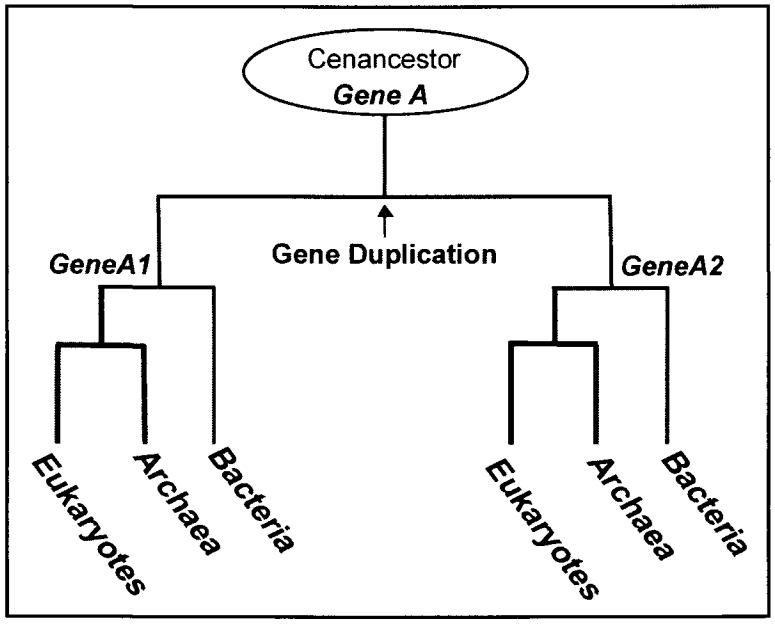

Figure 3. Conceptual rooting of the universal tree using paralogous genes. Gene A was duplicated in the cenancestor such that all extant organisms have paralogous copies, gene A1 and gene A2. The two genes are sufficiently similar to allow for the construction of reciprocally rooted trees thus rooting 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.[16] While the archaebacterium, *Pyrococcus abyssi*, was recently shown to have a bacteria-like origin of DNA replication, most of its replication enzymes are eukaryote-like.[17,18] Archaeal DNA scaffolding proteins are remarkably similar to eukaryotic histones.[19] Eukaryotes and the Archaea have similar transcriptional proteins, such as multi-subunit DNA-dependent RNA polymerases,[20] as well as sharing translation initiation factors not found in the Bacteria.[21,22] 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[23] aligned amino acids from five conserved regions shared by the elongation factors (EF) Tu/1α and EF-G/2 genes of the archaebacterium, *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[24] developed composite trees based on duplicated ATPase genes where the V-type A and V-type B occurs in Archaea and eukaryotes and the F_0F_1-type β and F_0F_1-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[25,26] and carbamoylphosphate synthetase[27] 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.[28-30] Philippe and coworkers[31,32] 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.[33] 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[34-36] 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*[37] and *Enterococcus hiraea*,[38] while a bacterial F_1- ATPase β subunit gene was found in the Archaea, *Methanosacrina barkeri*.[39] Consequently, Forterre and coworkers[40] suggested that the ATPase subunit gene family had not been fully determined, and that other paralogous family members might be discovered Hilario and Gogarten[41] 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.[42]

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.[43] In 1995 Golding and Gupta[44] 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[45] and Roger and Brown[46] 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[47] found that in the 34 universal protein trees they constructed, AE, AB and BE clusters occurred in the phylogenies 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.[13] 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.[48]

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.[49]

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.[50] 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.[51] Additional support for this view is the conservation of genomic context for translation-associated genes in bacteria.[52]

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.[48] Orthologous genes to eukaryotic glutaminyl-tRNA synthetase occur in many proteobacteria and *D. radiodurans* but not in other Bacteria or the Archaea.[56] Archaea and some bacteria, Spirochaetes, share novel type of lysyl-tRNA synthetases[57] and phenylalanyl-tRNA synthetases.[55,58,59]

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.[60-62] The genes are also found in the Archaea and the bacterial spirochaete, *Borrelia burgdorferi*. However, the mevalonate pathway is absent from the completely sequenced genome of a closely related Spirochaete, *Treponema pallidum*, and the Archaea have likely substituted an analogous protein for at least one enzyme in the pathway.[63] 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.[64,65]

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.[66,67] 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.[68] However, there is a need to critically evaluate methods for detecting HGT, which in some cases, can lead to overestimates of its occurrence.[36,69]

Reports of HGT without supporting phylogenetic analyses should be carefully scrutinized. Comparative studies based on BLAST[70] analyses have concluded that HGT has extensively occurred between Archaea and Bacteria. Koonin and coworkers[71] found that 44 % of the gene products of the archaebacterium, *Methanococcus jannaschii* were more similar to bacterial over

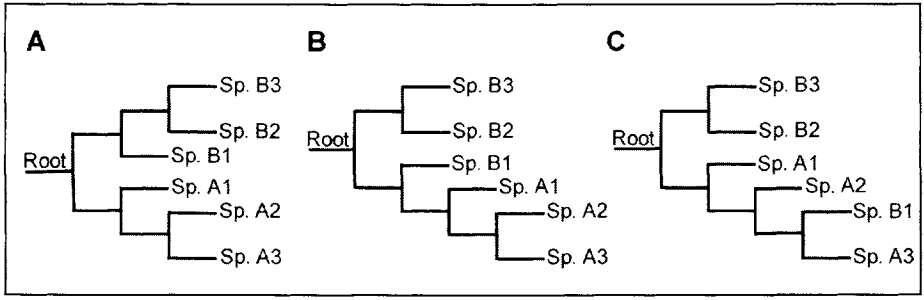

Figure 4. Detection of horizontal gene transfer (HGT) from phylogeny. Hypothetical protein trees for three bacterial species (B1-B3) and three archaeal species (A1-A3). A) The true rooting of the tree postulates a split between the Archaea and Bacteria, which results in two monophyletic clusters. B) 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[72] 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.[36,73] 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.[72,73] 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.[74]

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.[75] 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.[76-78]

In our study,[78] we examined all 28 cases where the IHGSC[75] had verified the presence of the gene in the human genome by PCR. BLAST[70] 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.[79] However, intragenomic base composition can be highly variable between chromosomal regions which could lead to over estimates in the number of transferred genes.[80,81] 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.[64] 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.[56]

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 *trans*-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.[13] 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.[82]

Among the three possible universal tree scenarios, only trees with the AE clustering depict, even if occasionally, all three Domains to be monophyletic simultaneously.[13] 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 *trans*-Domain HGT previously unnoticed in higher eukaryotes.[83]

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,[84] glutamate dehydrogenase[85] and HSP70[86] 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.[87] Catalase-peroxidase genes appear to have been exchanged between Archaea and pathogenic proteobacteria.[88] Two component signal transduction systems in the Archaea as well as fungi and slime molds were likely acquired from the Bacteria.[89] 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.[53,90] 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[91] 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 cyanobacterium, the bacterial progenitor of plastids.[92] 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.[93-97] 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.[83,98]

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.[15,99,100] 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.[101]

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[102,103] although overall neighborhoods of genes on the chromosome might be preserved because of functional and regulatory consequences.[59]

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[104] and eukaryotic protists.[105]

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.[106] 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.[107] 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 (Table 1). For example, the Clusters of Orthologous Groups of proteins (COGs) database (http://www.ncbi.nlm.nih.gov/COG/xindex.html) reports for 34 complete genomes, a total of 78 completely conserved proteins.[108] 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 (Fig. 5). 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.[54] 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[107]

	Cellular Function	Protein Name	Number of Amino Acids[a]	Support for Domain Monophyly[b]		
				Archaea	Bacteria	Eucary
1	translation	alanyl-tRNA synthetase	502	100	–	100
2		aspartyl-tRNA synthetase[c]	249	–	100	100
3		glutamyl-tRNA synthetase[c]	188	50 (–)	100	100
4		histidyl-tRNA synthetase	166	–	–	100 (93)
5		isoleucyl-tRNA synthetase	552	–	–	–
6		leucyl-tRNA synthetase[c]	358	–	100	100
7		methionyl-tRNA synthetase	306	–	–	99
8		phenylalanyl-tRNA synthetase b subunit	177	–	–	100
9		threonyl-tRNA synthetase	305	–	– (34)	100
10		valyl-tRNA synthetase	538	–	–	100
11		initiation factor 2[c]	337	–	100	100
12		elongation factor G[c]	536	64(87)	100	100
13		elongation factor Tu[c]	340	– (42)	100	100
14		ribosomal protein L2[c]	192	46(–)	100	100
15		ribosomal protein S5[c]	131	46(19)	100	100(99)
16		ribosomal protein S8[c]	118	–	100	100
17		ribosomal protein S11[c]	110	–	100	100
18		aminopeptidase P	95	–	–	–
19	transcription	DNA-directed RNA polymerase b chain[c]	537	99(78)	100	100
20	DNA replication	DNA topoisomerase I[c]	236	–	100	100
21		DNA polymerase III subunit[c]	194	46(49)	100	100(95)
22	metabolism	signal recognition particle protein[c]	298	71(39)	100	100
23		rRNA dimethylase	126	–	–	100(98)
		full alignment length[d]	6591			
		truncated alignment length[e]	3824			

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

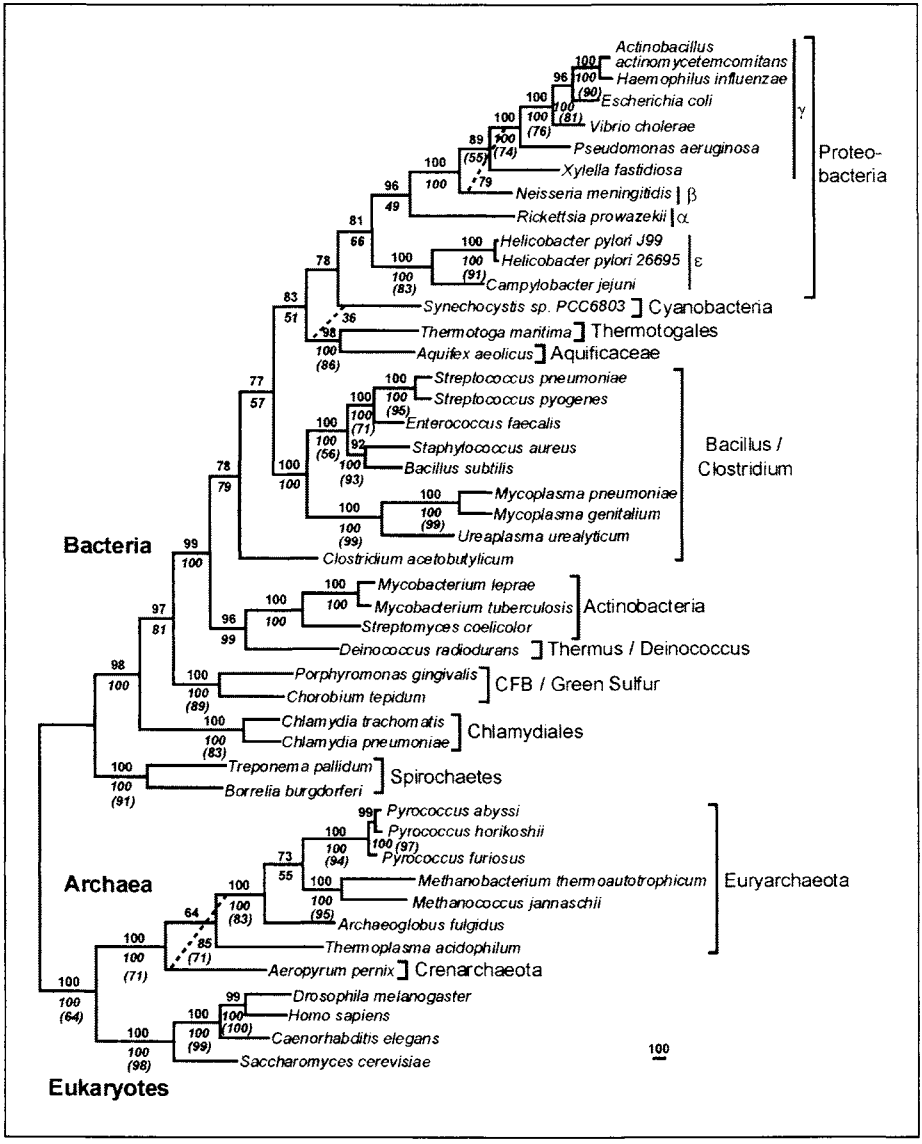

Figure 5. Universal tree based on 23 combined protein datasets.[107] Minimal length maximum parsimony universal tree based on 23 combined protein datasets is shown. Spirochaetes are placed as the lowest branching Bacteria. Numbers along the branches show the percent occurrence of nodes in 50% or greater of 1000 bootstrap replicates of maximum parsimony[122] (plain text) and neighbor joining[123] (italicized text) analyses or 1000 quartet puzzling steps of maximum likelihood[124] 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.

otes or the Archaea. When nine putatively horizontally transferred proteins were removed from the combined protein dataset, the truncated combined protein alignment was reduced to 3824 amino acids (Table 1).

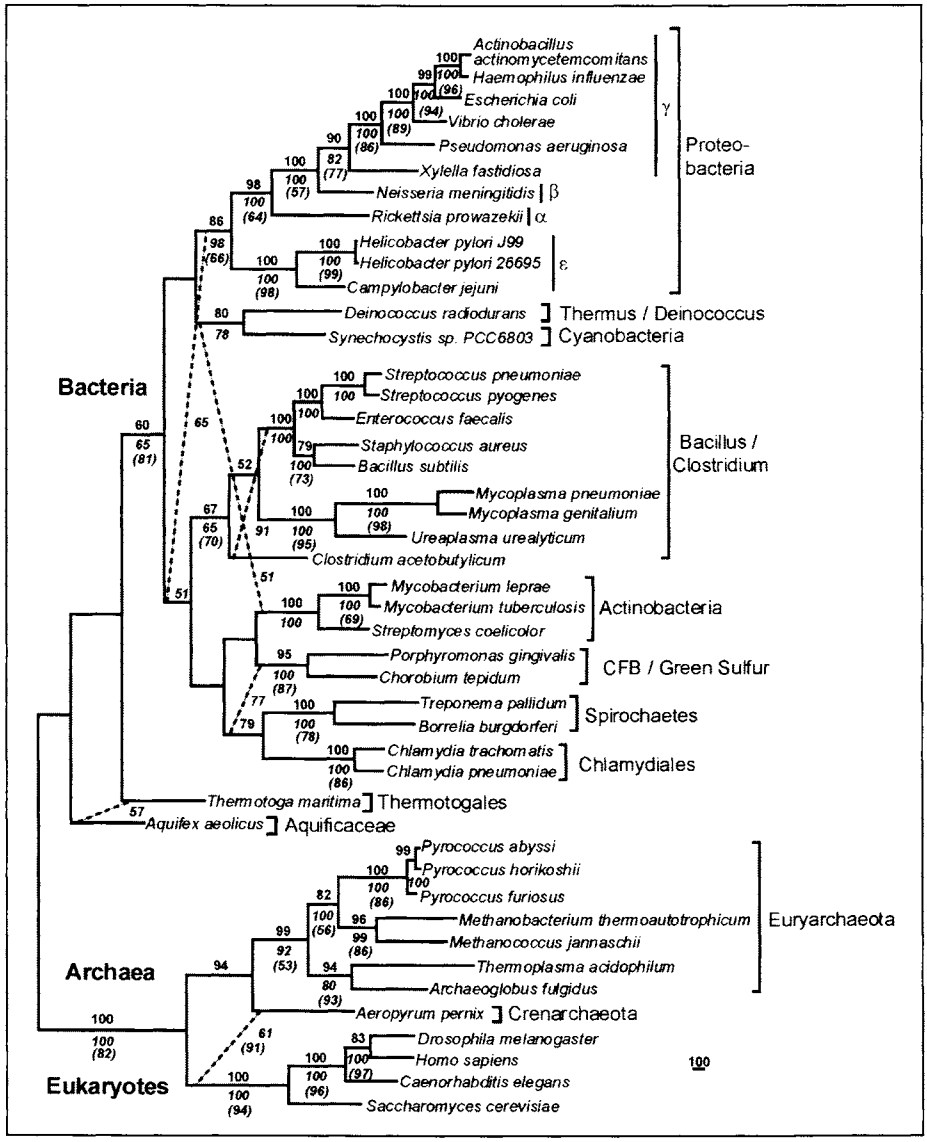

Figure 6. Universal tree based on 14 combined protein datasets. Minimal length maximum parsimony universal tree based on 14 proteins, with 9 horizontal gene transfer proteins removed, is shown. The tree shows Thermophiles as the basal group in Bacteria. Methods and labels are the same as Figure 5 and ref. 107. Figure adapted from ref. 107.

In contrast to the combined alignment of 23 proteins, phylogenetic trees based on the alignment of 14 nonHGT 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.[109] The clustering of Chlamydiales, CFB and Spirochaetes together is also novel relative to rRNA trees.[110] 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.[110-112] However, there are still many unresolved issues surrounding the "hot" origin of life hypothesis such as the maintenance of extracellular biochemical reactions[18] and the stability of RNA molecules at extreme temperatures.[113]

Genes found only in thermophilic Bacteria and Archaea are just as likely to be shared syplesiomorphies, 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 *Rickettsia prowazekii*, from which the endosymbiont progenitor of the mitochondria likely evolved.[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,[55,116] other studies based on wider collections of proteins support new relationships among bacterial groups.[102]

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.[41,117] Other radical positions discuss the emergence of eukaryotes from the complete fusion of genomes from an archaebacterium and bacterium (for a review see ref. 13). Martin and Müller[118] proposed a more stepwise progression to eukaryotes beginning with a hydrogen-dependent host, likely an archaebacterium, and a respiring bacterial symbiont. W.F. Doolittle[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.[121]

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. *Chlorobium tepidum, Enterococcus faecalis, Porphyromonas gingivalis*, and *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 *Actinobacillus actinomycetemcomitans* were obtained from the *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

1. Chatton E. Titres et Travauz Scientifiques. Setes Sattano Italy 1937.
2. Stanier RY, van Niel CB. The main outlines of bacterial classification. J Bacteriol 1941; 42:437-463.
3. Doolittle WF, Brown JR. Tempo, mode, the progenote, and the universal root. Proc Natl Acad Sci USA 1994; 91:6721-6728.
4. Stanier RY, van Niel CB. The concept of a bacterium. Arch Microbiol 1962; 42:17-35.
5. Stanier RY. Some aspects of the biology of cells and their possible evolutionary significance. Symp Soc Gen Microbiol 1970; 20:1-38.
6. Fox GE, Magrum LJ, Balch WE et al. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci USA 1977; 74: 4537-4541.
7. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci USA 1977; 51: 221-271.
8. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria and Eucarya. Proc Natl Acad Sci USA 1990; 87:4576-4579.
9. Stein JL, Simon MI. Archaeal ubiquity. Proc Natl Acad Sci USA 1996; 93: 6228-6230.
10. Danson MJ. Central metabolism of the Archaea. In: Kates M, Kushner DJ, Matheson AT, eds. The Biochemistry of Archaea (Archaebacteria). Amsterdam: Elsevier, 1993:1-24.
11. Kates M, Kushner DJ, Matheson AT. The Biochemistry of Archaea (Archaebacteria). Amsterdam: Elsevier, 1993.
12. Keeling PJ, Charlebois RL, Doolittle WF. Archaebacterial genomes: eubacterial form and eukaryotic content. Current Opinions in Genetics and Development 1994; 4:816-822.
13. Brown JR, Doolittle WF. Archaea and the prokaryote to eukaryotes transition. Microbiology and Molecular Biology Reviews 1997; 61:456-502.
14. Graham DE, Overbeek R, Olsen GJ et al. An archaeal genomic signature. Proc Natl Acad Sci USA 2000; 97: 3304-3308.
15. Snel B, Bork P, Huynen MA. Genome phylogeny based on gene content. Nature Genetics 1999; 21:108-110.
16. Edgell DR, Doolittle WF. Archaea and the origin(s) of DNA replication proteins. Cell 1997; 89:995-998.
17. Myllykallio H, Lopez P, López-Garcia P et al. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. Science 2000; 288:2212-2215.
18. Kelman Z. The replication origin of archaea is finally revealed. Trends in Biochem Sci 2000; 25:521-523.
19. Reeve JN, Sandman K, Daniels CJ. Archaeal histones, nucleosomes and transcription initiation. Cell 1997; 89:999-1002.
20. Langer D, Hain J, Thuriaux P et al. Transcription in Archaea: similarity to that in Eucarya. Proc Natl Acad Sci USA 1995; 92:5768-5772.
21. Olsen GJ, Woese CR. Archaeal genomics – an overview. Cell 1997; 89:991-994.
22. Kyrpides NC, Woese CR. Universally conserved translation initiation factors. Proc Natl Acad Sci USA 1998; 95:224-228.
23. Iwabe N, Kuma K-I, Hasegawa M et al. Evolutionary relationship of Archaea, Bacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 1989; 86:9355-9359.
24. Gogarten JP, Kibak H, Dittrich P et al. Evolution of the vacuolar H+-ATPase: Implications for the origin of eukaryotes. Proc Natl Acad Sci USA 1989; 86: 6661-6665.
25. Brown JR, Doolittle WF. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci USA 1995; 92:2441-2445.
26. Brown JR, Robb FT, Weiss R et al. Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. J Mol Evol 1997; 45:9-16.
27. Lawson FS, Charlebois RL, Dillon J-AR. Phylogenetic analysis of carbamoylphosphate synthetase genes: evolution involving multiple gene duplications, gene fusions, and insertions and deletions of surrounding sequences. Mol Biol Evol 1996; 13:970-977.

28. Lake JA. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature 1988; 331:184-186.
29. Rivera MC, Lake JA. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 1992; 257:74-76.
30. Baldauf SL, Palmer JD, Doolittle WF. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc Natl Acad Sci USA 1996; 93:7749-7754.
31. Lopez P, Forterre P, Philippe H. The root of the tree of life in the light of the covarion model. J Mol Evol 1999; 49:496-508.
32. Philippe H, Forterre P. The rooting of the universal tree of life is not reliable. J Mol Evol 1999; 49: 509-523.
33. Mazel D, Davies J. Antibiotic resistance in microbes. Cell Mol Sci 1999; 56:742-754.
34. de la Cruz F, Davies J. Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol 2000; 8:128-133.
35. Eisen JA. Horizontal gene transfer among microbial genomes: New insights from complete genome analysis. Curr Opin Genet Dev 2000; 10:606-611.
36. Kyrpides NC, Olsen GJ. Archaeal and bacterial hyperthermophiles: Horizontal gene exchange or common ancestry? Trends Genet 1999; 15:298-299.
37. Tsutsumi S, Denda K, Yokoyama K et al. Molecular cloning of genes encoding major subunits of a eubacterial V-type ATPase from Thermus thermophilus. Biochim Biophys Acta 1991; 1098:13-20.
38. Kakinuma Y, Igarishi K, Konishi K et al. Primary structure of the alpha-subunit of vacuolar-type Na+-ATPase in Enterococcus hirae, amplification of a 1000 bp fragment by polymerase chain reaction. FEBS Lett 1991; 292:64-68.
39. Sumi M, Sato MH, Denda K et al. A DNA fragment homologous to F1-ATPase beta-subunit amplified from genomic DNA of Methanosarcina barkeri: Indication of an archaebacterial F-type ATPase. FEBS Lett 1992; 314:207-210.
40. Forterre P, Benachenhou-Lahfa N, Confalonieri F et al. The nature of the last universal ancestor and the root of the tree of life, still open questions. Biosystems 1993; 28:15-32.
41. Hilario E, Gogarten JP. Horizontal transfer of ATPase genes — the tree of life becomes the net of life. BioSystems 1993; 31:111-119.
42. Gogarten JP, Hilario E, Olendzenski L. Gene duplications and horizontal transfer during early evolution. In: Roberts DM, Alderson G, Sharp P et al, eds. Evolution of Microbial Life. Society for General Microbiology Symposia 54. Cambridge: Cambridge University Press, 1996:267-292.
43. Smith MW, Feng D-F, Doolittle RF. Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem Sci 1992; 17:489-493.
44. Golding GB, Gupta RS. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol Biol Evol 1995; 12:1-6.
45. Gupta RS, Golding GB. The origin of the eukaryotic cell. Trends in Biochem Sci 1996; 21:166-171.
46. Roger AJ, Brown JR. A chimeric origin for eukaryotes reexamined. Trends Biochem Sci 1996; 21:370-371.
47. Feng D-F, Cho G, Doolittle WF. Determining divergence times with a protein clock: Update and reevaluation. Proc Natl Acad Sci USA 1997; 94:13028-13033.
48. Brown JR, Zhang J, Hodgson JE. A bacterial antibiotic resistance gene with eukaryotic origins. Curr Biol 1998; 8:R365-R367.
49. Koonin EV, Mushegian AR, Bork P. Nonorthologous gene displacement. Trends Genet 1996; 12:334-336.
50. Rivera MC, Jain R, Moore JE et al. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA 1998; 95:6239-6244.
51. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 1999; 96:3801-3806.
52. Lathe WC, Snel B, Bork P. Gene context conservation of a higher order than operons. Trends Biochem Sci 2000; 25:474-479.
53. Brown JR. Aminoacyl-tRNA synthetases: Evolution of a troubled family. In: Wiegel J, Adams M, eds. Thermophiles – the keys to molecular evolution and the origin of life. London: Taylor & Francis Group Ltd, 1998:217-230.
54. Wolf YI, Aravind L, Grishin NV et al. Evolution of aminoacyl-tRNA synthetases – analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. Genome Res 1999; 9:689-710.
55. Woese CR, Olsen GJ, Ibba M et al. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 2000; 64:202-236.
56. Brown JR, Doolittle WF. Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. J Mol Evol 1999; 49:485-95.

57. Ibba M, Morgan S, Curnow AW et al. A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. Science 1997; 278:1119-1122.
58. Teichmann SA, Mitchison G. Is there a phylogenetic signal in prokaryote proteins? J Mol Evol 1999; 49:98-107.
59. Brown JR. Genomic and phylogenetic perspectives on the evolution of prokaryotes. Systematic Biology 2001; 50:497-512.
60. Boucher Y, Doolittle WF. The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. Mol Microbiol 2000; 37:703-716.
61. Doolittle WF, Logsdon Jr JM. Archaeal genomics: Do archaea have a mixed heritage? Curr Biol 1998; 8:R209-R211.
62. Wilding EI, Brown JR, Bryant A et al. Identification, evolution and essentiality of the mevalonate pathway for isopentenyl diphosphate biosynthesis in Gram-positive cocci. J Bacteriol 2000; 182:4319-4327.
63. Smit A, Mushegian A. Biosynthesis of isoprenoids via mevalonate in Archaea: the lost pathway. Genome Res 2000; 10:1468-1484.
64. Lawrence JG. Selfish operons and speciation by gene transfer. Trends Microbiol 1997; 5:355-359.
65. Lawrence JG, Roth JR. Selfish operons – horizontal transfer may drive the evolution of gene clusters. Genetics 1996; 143:1843-1860.
66. Pennisi E. Genome data shake the tree of life. Science 1998; 280:672-674.
67. Pennisi E. Is it time to uproot the tree of life? Science 1999; 284:1305-1307.
68. Doolittle WF. Phylogentic classification and the universal tree. Science 1999; 284:2124-2128.
69. Kurland C. Something for everyone: Horizontal gene transfer in evolution. EMBO Reports 2000; 1:92-95.
70. Altschul SF, Madden TL, Schäffer AA et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389-3402.
71. Koonin EV, Mushegian AR, Galperin MY et al. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol Microbiol 1997; 25:619-637.
72. Nelson KE, Clayton RA, Gill SR et al. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 1999; 399:323-329.
73. Logsdon Jr JM, Faguy DM. Evolutionary genomics: Thermotoga heats up lateral gene transfer. Curr Biol 1999; 9:R747-R751.
74. Olendzenski L, Liu L, Zhaxybayeva O et al. Horizontal transfer of archaeal genes into the deinococcaceae: detection by molecular and computer-based approaches. J Mol Evol 2000; 51:587-599.
75. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409:860-921.
76. Roelofs J, van Haastert PJM. Genomics: Genes lost during evolution. Nature 2001; 411:1013-1014.
77. Salzburg SL, White O, Peterson J et al. Microbial genes in the human genome: lateral transfer or gene loss? Science 2001; 292:1903-1906.
78. Stanhope MJ, Lupas AN, Italia MJ et al. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature 2001; 411:940-944.
79. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature 2000; 405:299-304.
80. Lafay B, Lloyd AT, McLean MJ et al. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 1999; 27:1642-1649.
81. Guindon S, Perrière G. Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes. Mol Bio Evol 2001; 18:1838-1840.
82. Brinkman H, Philippe H. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 1999; 16:817-825.
83. Chihade J, Brown JR, Schimmel P et al. Detection of an intermediate stage of mitochondria genesis. Proc Natl Acad Sci USA 2000; 97:12153-12157.
84. Brown JR, Masuchi Y, Robb FT et al. Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J Mol Evol 1994; 38:566-576.
85. Benachenhou-Lahfa N, Forterre P, Labedan B. Evolution of glutamate dehydrogenase genes: Evidence for paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. J Mol Evol 1993; 36:335-346.
86. Gupta RS, Golding GB. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria and eukaryotes. J Mol Evol 1993; 37:573-582.
87. Forterre P, Bouthier de la Tour C, Philippe H et al. Reverse gyrase from thermophiles: probable transfer of a thermoadaptation trait from Archaea to Bacteria. Trends in Genet 2000; 16:152-154.

88. Faguy DM, Doolittle WF. Horizontal transfer of catalase-peroxidase genes between Archaea and pathogenic bacteria. Trends in Genet 2000; 16:196-197.
89. Koretke KK, Lupas AN, Warren PV et al. Evolution of two-component signal transduction. Mol Biol Evol 2000; 17:1956-1970.
90. Lamour V, Quevillon S, Diriong S et al. Evolution of the Glx-tRNA synthetase family: The glutaminyl enzyme as a case for horizontal gene transfer. Proc Natl Acad Sci USA 1994; 91:8670-8674.
91. Margulis L. Origin of eukaryotic cells. New Haven: Yale University Press, 1970.
92. Rujun T, William M. How many genes in Arabidopsis come from cyanobacteria? An estimate from 386 protein phylogenies. Trends in Genet 2001; 17:113-120.
93. Clark CG, Roger AJ. Direct evidence for secondary loss of mitochondria in Entamoeba histolytica. Proc Natl Acad Sci USA 1995; 92:6518-6521.
94. Germot A, Philippe H, Le Guyader H. Presence of a mitochondrial-type 70-kDa heat shock protein in Trichomonas vaginalis suggests a very early mitochondrial endosymbiosis in eukaryotes. Proc Natl Acad Sci USA 1996; 93:14614-14617.
95. Hashimoto T, Sánchez LB, Shirakura T et al. Secondary absence of mitochondria in Giardia lamblia and Trichomonas vaginalis revealed by valyl-tRNA synthetase phylogeny. Proc Natl Acad Sci USA 1998; 95:6860-6865.
96. Henze KA, Badr A, Wettern M et al. A nuclear gene of eubacterial origin in Euglena gracilis reflects cryptic endosymbioses during protist evolution. Proc Natl Acad Sci USA 1995; 92:9122-9126.
97. Keeling PJ, Doolittle WF. Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. Proc Natl Acad Sci USA 1997; 94:1270-1275.
98. Brown JR, Italia MJ, Douady C et al. Horizontal gene transfer and the universal tree of life. In: Syvanen M, Kado CI, eds. Horizontal Gene Transfer. 2nd eds. Academic Press, 2002: In press.
99. Huynen M, Snel B, Bork P. Lateral gene transfer, genome surveys and the phylogeny of prokaryotes. Technical Comments. Science 1999; 286:1443a.
100. Lin J, Gerstein M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res 2000; 10:808-818.
101. Doolittle WF. Lateral gene transfer, genome surveys and the phylogeny of prokaryotes. Technical Comments. Science 1999; 286:1443a.
102. Wolf YI, Rogozin IB, Grishin NV et al. Genome trees constrcuted using five different approaches suggest new major bacterial clades. BMC Evolutionary Biology 2001 1:8.
103. Wolf YI, Rogozin IB, Kondrashov AS et al. Genome alignment, evolution of prokaryotic genome organization,and prediction of gene function using genomic content. Genome Res 2001; 11:356-372.
104. Xiong J, Inoue K, Bauer CE. Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from Heliobacillus mobilis. Proc Natl Acad Sci USA 1998; 95:14851-14856
105. Baldauf SL, Roger AJ, Wenk-Siefert I et al. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 2000; 290:972-977.
106. Gribaldo S, Lumia V, Creti R et al. Discontinous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bacteriol 1999; 181:434-443.
107. Brown JR, Douady CJ, Italia MJ et al. Universal trees based on large combined protein sequence datasets. Nature Genetics 2001; 28: 281-285.
108. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science 1997; 278:631-637.
109. Shah HN, Gharbia SE, Collins MD. The Gram stain: a declining synapomorphy in an emerging evolutionary tree. Rev in Med Microbiol 1997; 8:103-100.
110. Olsen GJ, Woese CR, Overbeek R. The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 1994; 176:1-6.
111. Woese CR. Bacterial evolution. Microbiol Rev 1987; 51:221-271.
112. Pace NR. Origin of life—Facing up to the physical setting. Cell 1991; 65:531-533.
113. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. Science 1999; 283:220-221.
114. Kurland C, Andersson SGE. Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev 2000; 64:786-820.
115. Andersson SG, Zomorodipour A, Andersson JO et al. The genome sequence of Rickettsia prowzekii and the origin of mitochondria. Nature 1998; 396: 133-140.
116. Hansmann S, Martin W. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int J Syst Evol Microbiol 2000; 50:1655-1663.

117. Woese CR. The universal ancestor. Proc Natl Acad Sci USA 1998; 51:221-271.
118. Martin W, Müller M. The hydrogen hypothesis for the first eukaryote. Nature 1998; 392:37-41.
119. Doolittle WF. You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends in Genet 1998; 14:307-311.
120. Keeling PJ, McFadden GI. Origins of microsporidia. Trends in Microbiol 1998; 6:19-23.
121. Brown JR. Ancient horizontal gene transfer. Nature Rev Gen 2003; 4:121-132.
122. Swofford DL. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b5. Sunderland: Sinauer Associates, 1999.
123. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author: http://evolution.genetics.washington.edu/phylip.html, Seattle: University of Washington, Seattle. 2000.
124. Strimmer K, von Haeseler A. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. Mol Biol Evol 1996; 13:964-969.