Space Sci Rev (2008) 135: 115–131
DOI 10.1007/s11214-007-9253-8

Gene Transfer and the Reconstruction of Life's Early History from Genomic Data

J. Peter Gogarten · Gregory Fournier · Olga Zhaxybayeva

Received: 22 December 2006 / Accepted: 30 July 2007 / Published online: 5 October 2007

Abstract The metaphor of the unique and strictly bifurcating tree of life, suggested by Charles Darwin, needs to be replaced (or at least amended) to reflect and include processes that lead to the merging of and communication between independent lines of descent. Gene histories include and reflect processes such as gene transfer, symbioses and lineage fusion. No single molecule can serve as a proxy for the tree of life. Individual gene histories can be reconstructed from the growing molecular databases containing sequence and structural information. With some simplifications these gene histories can be represented by furcating trees; however, merging these gene histories into web-like organismal histories, including the transfer of metabolic pathways and cell biological innovations from now-extinct lineages, has yet to be accomplished. Because of these difficulties in interpreting the record retained in molecular sequences, correlations with biochemical fossils and with the geological record need to be interpreted with caution. Advances to detect and pinpoint transfer events promise to untangle at least a few of the intertwined histories of individual genes within organisms and trace them to the organismal ancestors. Furthermore, analysis of the shape of molecular phylogenetic trees may point towards organismal radiations that might reflect early mass extinction events that occurred on a planetary scale.

Keywords Tree of life · Horizontal gene transfer · Late heavy bombardment · Coalescence

1 Trees, Forests, Webs: How to Depict Evolutionary History

1.1 Darwin's Coral of Life

Bifurcating trees have long been used to depict evolutionary history. The earliest depictions were family trees, tracing the ancestry of individuals. It is noteworthy that these family

J.P. Gogarten (✉) · G. Fournier
Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269-3125, USA
e-mail: gogarten@uconn.edu

O. Zhaxybayeva
Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS B3H 1X5, Canada

trees or pedigrees often have their root in the extant individual, and that the tree grows more branches as one moves back into the past (two parents, four grandparents, eight great-grandparents, etc.). In contrast, species trees trace back the history of groups of organisms to common ancestral groups. For example, all animals are thought to have evolved from a single-celled choanoflagellate ancestor (Lang et al. 2002; Philippe et al. 2004). Lamarck (1815), in his work on the classification of invertebrates, was the first to consider and depict species evolution as a bifurcating, tree-like process. Darwin recognized that species evolve, with natural and sexual selection causing a parent species to diverge into two different new species. As a logical consequence, he concluded that all living organisms could be traced back to a single ancestor (Darwin 1859). Darwin frequently used the term "tree of life", but he also pointed out that the term "coral of life" would be more appropriate. Like the tree of life, the base of the branches of coral is made out of extinct, dead organisms, whereas in botanical trees living cells are found throughout the tree (Darwin 1836–1844). The metaphor of a coral of life is also appropriate because in many corals the thin layer of living organisms sits on the massive, richly connected skeletons formed by their ancestors (compare Fig. 1).

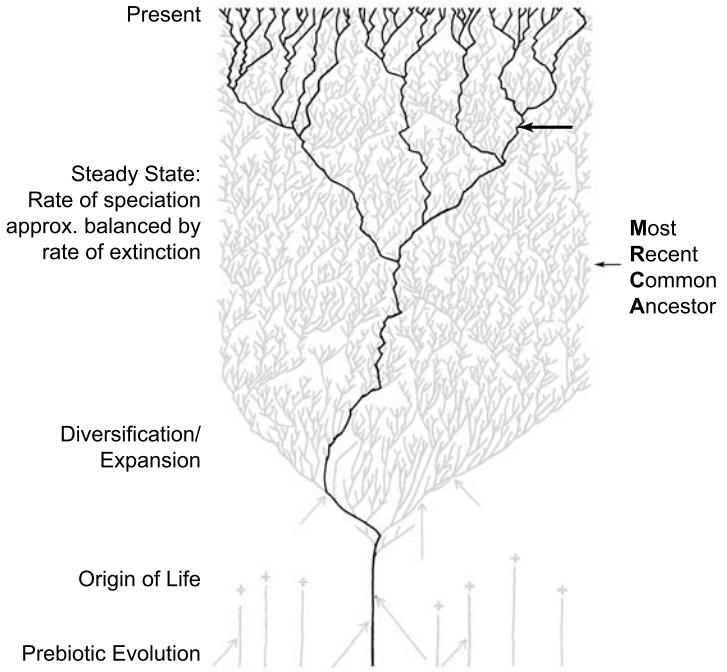

Fig. 1 Schematic representation of the "coral of life". In his notebook, Charles Darwin suggested that the term "coral of life" might be preferable to the term tree of life (Darwin 1836–1844), because the layer of living, extant organisms rests on dead branches. In this diagram lineages with extant representatives are drawn in black, whereas extinct lineages are given in gray. The depicted scenario assumes that life originated from one or several prebiotic chemical processes, and that after its origin life adapted to different ecological niches available on the early Earth. Later on evolution represents an approximate balance between extinction and speciation; the evolution of extant lineages can be described by a coalescence process (Zhaxybayeva and Gogarten 2004). Note that the most recent common ancestor (MRCA) of all organisms only came into existence some time after the origin of life, and that many other, now-extinct lineages existed at the same time as well. The *horizontal black arrow* exemplifies that some genes that evolved early in life's history might have been transferred from now extinct lineages into extant lineages

1.2 Intertwined Trees and Coalescence

1.2.1 Exchange Groups, Species

The biological species concept describes species as groups of organism that can create offspring (Mayr 1942). As a consequence of recombination, within a species molecular phylogenies will not be congruent, and innovations can be shared within the species. In case of prokaryotes, procreation is independent of recombination, and genetic exchange is not limited within a species. Species boundaries, especially within prokaryotes, are fuzzy because some genes transfer across them (Gogarten and Townsend 2005; Hanage et al. 2005). This finding is not restricted to prokaryotes (Arnold 2006), but among eukaryotes these processes are mainly observed in recently diverged species (e.g., Grant et al. 2004).

The exchange of genetic material between divergent organisms greatly accelerates evolution (Jain et al. 2003). Horizontal gene transfer (HGT) can transfer an invention or an improvement made in one part of the tree of life to other lineages. Microorganisms need not reinvent a metabolic pathway, they can acquire it from other organism. It therefore is not surprising that genes encoding enzymes involved in metabolism are frequently found transferred between very divergent organisms (see Sect. 3.2.1).

However, HGT has not been so rampant as to create a continuum of phenotypes. A strain of *Escherichia coli* is clearly recognized by its pheno- and genotype as belonging to this species, which in turn is clearly placed within the gamma proteobacteria, which in turn are clearly identified as bacteria. This clear identification as *E.coli* is made even though the comparison of genomes of three *E. coli* strains revealed that less than 40% of the common shared gene pool was represented in all three genomes (Welch et al. 2002). Despite this enormous within-species diversity with respect to genome content, the measure traditionally used to define species boundaries in prokaryotes (70% DNA reassociation after melting the DNA mixture from two organisms) is in good agreement with percent SSU rRNA identity (about 98%) (Rossello-Mora and Amann 2001) and average nucleotide identity in those genes that are present in both organisms (more than 94%) (Konstantinidis and Tiedje 2005).

Currently, three extreme models are discussed to explain the cohesion within species (Gevers et al. 2005), and probably each of these is close to reality in some instances: First, in species with little recombination an advantageous mutation that spreads through a population will carry the whole genome along, thereby leading to populations that only recently diverged from their Most Recent Common Ancestor (MRCA); second, in species with a high rate of recombination, the biological species concept might apply, provided that within-species recombination is more frequent than gene flow between species; and third, species might appear coherent, because they separated a long time ago, and the intermediate forms went extinct.

1.2.2 Pedigrees and Species Trees

The trees of family history depicting the ancestry of specific individuals are all deeply intertwined through recombination and sexual procreation. Entwined together, the family trees of all the members of a species form the lineage depicting the descent of the entire species. Somewhere within this lineage there exists an individual from which all living members of a species can claim descent. In sexual interbreeding populations, the time to the most recent of these common ancestor is surprisingly short, because the total number of ancestors for each individual grows geometrically (at least initially) as one moves back in time. For example, the most recent common ancestor of all humans was estimated to have lived just a few

thousand years ago, even under the assumption of mating occurring mainly in geographically determined subpopulations (Rohde et al. 2004). Note that the most recent common ancestor defined in the sense of a pedigree has not necessarily contributed any genes to all its descendants' genomes.

1.2.3 Gene Trees and Species Trees

In some regards things are simplified, if one adopts a gene-centered view of evolution in which organisms are considered the vessels constructed by the genes in order to be propagated into future generations (Dawkins 1976). Gene trees, in contrast to pedigrees, bifurcate in the same way as species trees—i.e., the individual gene usually is derived from only one immediate ancestor and it can have multiple descendents. Genomic recombination, gene transfer and lineage sorting can cause differences between trees defined by speciation and trees defined by the descent of a particular gene (Page and Holmes 1998; Felsenstein 2003). The most recent common ancestor of two individuals belonging to different species or of two alleles present in a population might have lived some time before the two species actually separated. In small populations with frequent recombination between individuals this time difference usually can be ignored because the time required for lineage sorting is negligible compared to the time separating two speciation events. However, in case of large populations fixation due to genetic drift requires more time on average; in populations with infrequent recombination between genomes, the time required for lineage sorting might no longer be negligible in a geographically dispersed population. Both of these factors, large population size and low recombination frequency, are common in prokaryotic populations, where procreation is not linked to recombination and effective population sizes are estimated to be larger than 10^8 individuals (Lynch and Conery 2003; Lynch 2006). Longer time intervals for lineage sorting also result when different coexisting alleles provide distinct adaptive advantages.

The inverse problem—i.e. the gene ancestor being more recent than the species ancestor—occurs perhaps more frequently. Species boundaries are not impermeable, and often organisms from recently diverged species can and do interbreed (Arnold 2006). Thus, the most recent common ancestor of two individuals from different species might be more recent or more ancient than the speciation event that separates the two species. Although gene transfer across species boundaries occurs perhaps more frequently in prokaryotes than in eukaryotes, the differences between gene and species trees that are due to lineage sorting and gene flow occur in eukaryotes as well as prokaryotes.

1.2.4 Most Recent Common Ancestors in Gene and Species Trees

In sexual eukaryotic populations, whole genomes line up and can undergo homologous recombination. The linkage between different genes breaks down and each individual gene (or gene fragment) has its individual history. Uniparentally inherited genes in the recombining human population provide a useful illustration of individual gene histories being different from one another. The non-recombining part of the Y-chromosome coalesces to a common ancestor that lived only about 50,000 years ago (Thomson et al. 2000; Underhill et al. 2000), while the mitochondrial genomes in humans (inherited via the female lineage) have a common ancestor that existed about 200,000 years before present (Cann et al. 1987; Vigilant et al. 1991). Y-chromosome Adam never met mitochondrial Eve, and at the time the ancestral genes existed, many other mitochondrial genomes and Y-chromosomes co-existed in the same population. However, these other genes did not make it into today's human population. Y chromosome Adam and mitochondrial Eve were not alone, and many of their

human contemporaries likely contributed other genes to the now existing gene pool; however, because of recombination, these other gene lineages cannot be traced as effectively as in the case of mitochondria and Y chromosomes.

1.2.5 Prokaryotic Evolution and the Trees of Life

In case of prokaryotes the situation appears even more complicated. Recombination is not restricted to members of the same species, since genes can be transferred between divergent organisms. Furthermore, the evolutionary lineages of prokaryotic genomes are surrounded by a halo of mobile genetic elements that only sometimes reside in genomes, but perhaps more often reside in phages, viruses and plasmids. Many genes found in microbial genomes were acquired from this cloud of transitory genes, and often these genes do not persist in the genome for long periods of time (Daubin et al. 2003; Gogarten and Townsend 2005; Hsiao et al. 2005).

This pool of horizontally transferred genes is characterized by nucleotide (high A and T content) and codon usage biases. These biases might reflect the mutation bias (mutations from GC pairs to AT pairs occur more frequently than vice versa), and indicate that these genes are under weak purifying selection only. Most of the genes transferred between prokaryotes belong to this transitory, weakly selected pool of genes; however, all types of genes, including ribosomal RNA and protein coding genes, are known to have been transferred (Gogarten et al. 2002; Lawrence and Hendrickson 2003) (see Sect. 3 for more discussion).

In view of gene transfer between divergent species, it is not surprising to find that different genes have different histories. For example, the tyrosyl-tRNA synthetases of animals and fungi group with their haloarchaeal homologs (Huang et al. 2005), the ATP synthases of the Deinococcaceae (a group of bacteria, with bacterial cell wall, bacterial membranes and bacterial ribosomes) group with homologs from archaea (Olendzenski et al. 2000; Lapierre et al. 2006). Clearly, every gene has its own phylogeny. Ignoring intragenic recombination, the history of any individual gene can be described by a bifurcating tree. Yet these trees are different from one another, and in particular, the most recent common ancestors of individual genes, even if they are present in all extant organisms, might have existed at different times and in different lineages (see Fig. 1 and Zhaxybayeva and Gogarten (2004)). Relationships between organismal lineages may generally take the form of a bifurcating tree. However, organismal evolution certainly will include some reticulations. For example, eukaryogenesis (the emergence of the eukaryotic cell) involved at least the symbiosis between an archaea related host cell and a bacterial endosymbiont that became the ancestor of today's mitochondria (Gogarten et al. 1989; Margulis 1995; Martin and Koonin 2006). More importantly, while we might be able to define an organismal most recent common ancestor, we need to remain aware that most recent common ancestors of individual genes probably did not exist together inside the organismal most recent common ancestor.

The extension of population genetics principles to the evolution of microbial species does not depend on horizontal exchange during the early evolution of life being more frequent than today (Woese 1998): even the low rates of exchange between divergent organisms observed today are sufficient justification. However, if rates of genetic exchange were higher than they are today, this would argue even more strongly for the application of population genetics to the early evolution of life.

1.3 The Shape of the Tree of Life

1.3.1 Forces and Processes that Lead to Bush-Like Phylogenies

Many evolutionary and analytical processes shape phylogenetic trees. This shape can be described through "lineages through time" plots that give the number of lineages existing over time (Raup 1985). In many instances it was observed that the number of lineages increases in an exponential fashion over time (Martin et al. 2004). A rapid succession of speciation events leads to a bush-like phylogeny, also known by the term radiation (Rokas and Carroll 2006). Other processes that lead to phylogenies with similar appearance are extinction events and the under-sampling of present species. Species radiations can follow a large ecological change that provides for many empty ecological niches that can be occupied by subpopulations that subsequently evolve into distinct species. These ecological changes can be caused by biology [e.g., disruptions due to invasive species (Gurevitch and Padilla 2004)] or by geological or astronomical processes [e.g., mass extinctions caused by meteorite impacts (Raup 1989; Gogarten-Boekels et al. 1995; Nisbet and Sleep 2001)].

1.3.2 Coalescence and Long, Empty, Basal Branches

Coalescence is the process of tracing lineages backwards in time to their common ancestors (Felsenstein 2003). Every two extant lineages coalesce to their most recent common ancestor. Eventually, all lineages coalesce to one lineage. The shape of the resulting tree depends on whether all lineages or only extant ones are considered. For example, animal phylogenies frequently include extinct organisms known from the fossil record. In contrast, molecular phylogenies are based on molecules from extant species only, with the possible exception of a few species that went extinct recently, but whose DNA has survived to the present day [e.g., the wooly mammoth or the Neanderthal humans (Gibbons 2005; Willerslev and Cooper 2005; Donoghue and Spigelman 2006)].

Considering only lineages with extant representatives greatly impacts the shape of the resulting phylogeny. A simple steady-state model in which speciation is balanced by extinction (compare Fig. 1) results in a faster than exponential growth in lineages through time plots (Zhaxybayeva and Gogarten 2004). The branching pattern in the phylogenies derived from this simple model is described by the Kingman coalescence (Felsenstein 2003), in which the coalescence of the last two lineages to their common ancestor occupies on average about half the time of the total coalescence process. Under the assumption of a steady state between extinction and speciation, for every group defined by shared ancestry, on average the two deepest branches are expected to cover half the time the group is in existence (Zhaxybayeva and Gogarten 2004). In the case of some molecules and for some features of the trees of life this expectation is met, for example, the ancestors of the archaeal and bacterial domain frequently are connected to their common ancestor by long branches (Gogarten-Boekels et al. 1995). Deviations from this expectation point towards processes that deviate from the simple steady-state model (see Sect. 1.3.3).

1.3.3 The Bacterial Radiation

One of the deviations from the expectation of long basal branches occurs in the early evolution of the bacterial domain. Bergey's manual, the main reference for bacterial and archaeal taxonomy recognizes 24 different bacterial phyla (Garrity et al. 2004) (see Fig. 2). Many

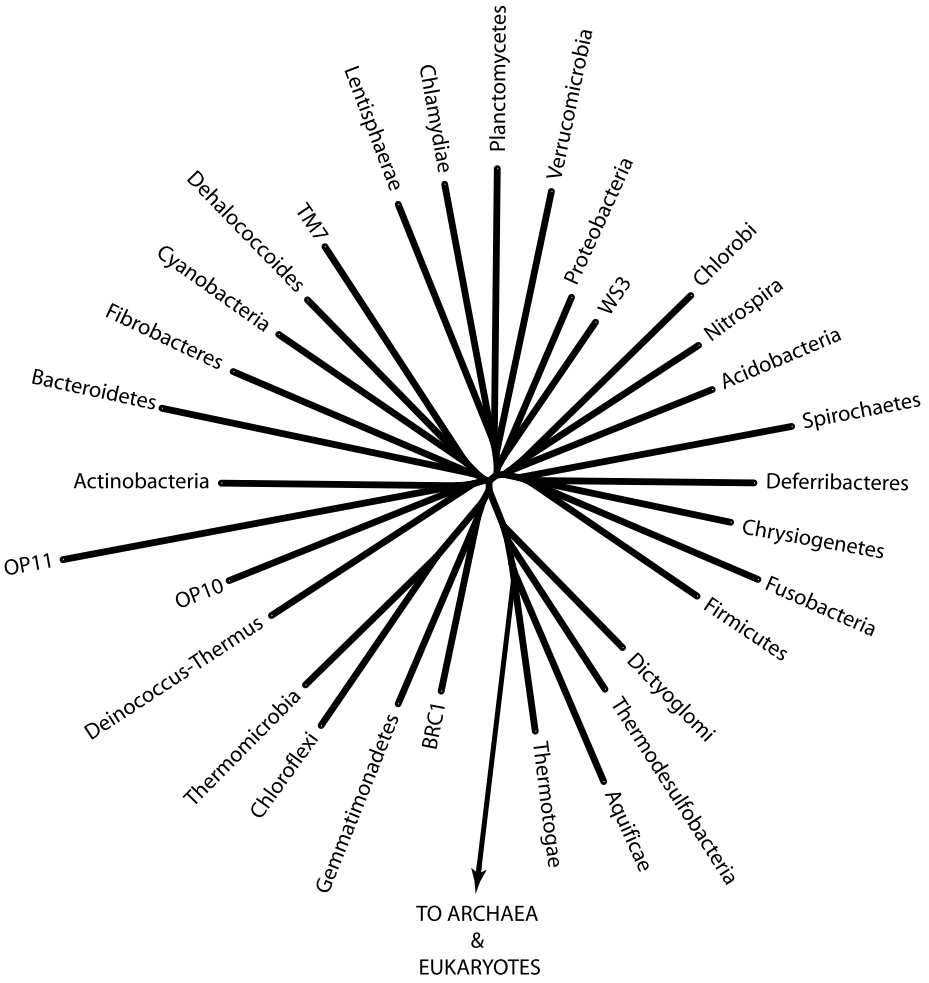

Fig. 2 Unrooted phylogenetic tree for 31 bacterial phyla listed in the Ribosomal Databank (RDP) version 9.44 (Cole et al. 2005). Each phylum is represented by one branch. The alignment and tree were obtained from the RDP website (http://rdp.cme.msu.edu/). Note the deviation in the tree shape from "long branches leading to MRCA" observed under simple birth-death models (Zhaxybayeva and Gogarten 2004). The shape of the shown tree suggest an actual radiation (a rapid succession of bifurcations) at the base of the bacterial domain

additional bacterial phyla are recognized based on amplification of genes encoding ribosomal RNA isolated from the environment (e.g., Ley et al. 2006). Many of these bacterial phyla include only organisms that have not been cultured at present (Schloss and Handelsman 2004). All of these ribosomal RNA defined phyla (see Sect. 2) diverged over a very small phylogenetic distance very close to the root of the bacterial domain (Fig. 2). At present the branching order between the bacterial phyla has to be considered as unresolved. The reasons for this are two-fold: First, the outgroup that can be used to place the domain ancestor relative to the bacterial phyla (either archaea or the eukaryotic nucleocytoplasm) is connected to the bacteria by a very long branch; the placement of the end of this branch inside the bacterial domain is difficult, and might be subject to artifacts. Second,

the branches separating the different phyla are short; even omitting the outgroup the relationships cannot be resolved with confidence. These findings suggest that a real radiation occurred at the base of the bacterial domain. One reason for this radiation to occur might have been a mass extinction that was triggered by the late heavy bombardment (Gomes et al. 2005). Only organisms that before the bombardment had adapted to deep-surface environments had a chance to survive this violent episode in Earth's history. After the end of the late heavy bombardment these survivors would have found many ecological niches closer to Earth's surface into which to adapt (Raup 1989; Gogarten-Boekels et al. 1995; Nisbet and Sleep 2001).

2 The Tree of Life According to Ribosomal RNA

2.1 Attempts to Create a Prokaryotic Taxonomy

Traditionally, single-celled organisms without a nucleus were placed into a group called monera (Haeckel 1866) or prokaryotes (Stanier and Van Niel 1962). Classification of single-celled organisms within this group was initially based on morphology (e.g., cell shape), physiology and biochemistry (e.g., fermentation type, and temperature ranges for growth) [see more on the history of prokaryotic taxonomy in Rossello-Mora and Amann (2001); Olendzenski et al. (2004); Sapp (2005)].

The study of ribosomal RNA (rRNA) molecules by Carl Woese, George Fox and colleagues (Woese and Fox 1977) launched a new era in approaches to phylogeny and provided a measure by which all organisms could be compared. rRNA molecules are ubiquitous in distribution, and perform the same function in all organisms. Furthermore, rRNAs contain both highly conserved and highly variable regions that can be compared between organisms of very different degrees of relationship (Woese 1987). The ability to amplify rDNA (rRNA encoding DNA) outside the living organism allows analyses not only from cultured organisms, but also from environmental samples (e.g. Ley et al. 2006; Sogin et al. 2006). With the isolation of environmental rDNA sequences that did not match any cultured organisms came the realization that the diversity of the prokaryotes had been vastly underestimated (Schloss and Handelsman 2004; Sogin et al. 2006). It is generally believed that, depending on environment sampled, only between 0.001% and 15% of existing prokaryotic diversity has been cultured (Amann et al. 1995). Sequencing of rDNA has become a standard procedure to characterize cultures and environmental samples. On December 6, 2006, more than 286,000 aligned sequences were available in the ribosomal RNA databank (Cole et al. 2005); this wealth of sequence information facilitates phylogenetic placement of organisms characterized by only small stretches of rDNA sequences (Sogin et al. 2006).

2.2 The Ribosomal RNA Based Tree of Life

Ribosomal RNA (rRNA) has been used to reconstruct the tree of life. Figure 3 summarizes some of the features frequently associated with the rRNA phylogeny. The ribosomal RNA-based tree of life clearly distinguishes three domains: Bacteria, Archaea, and Eukaryotes (Woese et al. 1990).

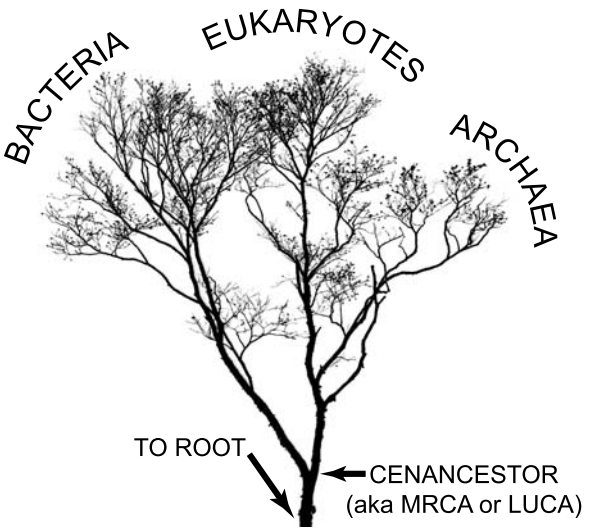

Fig. 3 Diagram of the small subunit ribosomal RNA-based tree of life. According to this tree, all life can be divided into three monophyletic groups, or domains (Bacteria, Archaea and Eukaryotes). The tree is strictly bifurcating, shows no reticulations, and only extant lineages are depicted (since 16S rRNA genes are sequenced from contemporary organisms). The tree is based on a single molecular phylogeny. All lineages on the tree trace back to a most recent common ancestor (MRCA). (The tree used as a backbone was photographed in the Point Pleasant Park in Halifax, Nova Scotia)

2.2.1 The Root of the Tree of Life

The tree of life calculated from rRNA sequences has to be considered unrooted. The placement of the root in the branch leading to the bacteria was not determined using rRNA, rather the placement of the root is based on the study of other evolutionary constrained molecules (ATPase, elongation factors, signal recognition particles) that underwent gene duplications early in life's history (Gogarten et al. 1989; Iwabe et al. 1989; Gogarten and Taiz 1992; Brown and Doolittle 1995; Gribaldo and Cammarano 1998); an overview on the current state of the sometimes controversial and ongoing debate on where to place the root in the tree of life is included in Zhaxybayeva et al. (2005).

The root of a molecular phylogeny refers to the deepest (oldest) bifurcation in the tree (and not to the bottom of the trunk). The organism represented by this deepest bifurcation is also known as the Most Recent Common Ancestor (MRCA) of all life, as cenancestor (Fitch and Upper 1987), or as the Last Universal Common Ancestor (LUCA). This organism existed at a rather long distance from the origin of life. Due to gene transfer, this organism—or population of organisms—almost certainly did not harbor all of the MRCAs of the different molecular systems found in all of today's organisms (see Sect. 1.2.4); however, this organism probably possessed membranes used in chemiosmotic coupling (Gogarten and Taiz 1992; Pereto et al. 2004), its genes were arranged on one or more chromosomes, and it used a complete genetic code encoding 20 amino acids (Anantharaman et al. 2002; Delaye et al. 2005).

2.2.2 Ribosomal RNA and Horizontal Gene Transfer

Like other molecular phylogenies, the small subunit rRNA-based tree of life is depicted as a strictly bifurcating tree (e.g., Woese et al. 1990). However, some organisms are known to harbor multiple, sometimes rather divergent, rRNA operons (Mylvaganam and Dennis 1992; Yap et al. 1999; Acinas et al. 2004) and there are indications that ribosomal RNA operons can be transferred between different species, and that subsequent to the transfer recombination between the divergent copies can occur (Dennis et al. 1998; Yap et al. 1999;

Gogarten et al. 2002; Boucher et al. 2004). Although these transfer and recombination events can make it difficult to identify organisms accurately (Morandi et al. 2005), most of these events took place between closely related organisms. Since homologous recombination requires short stretches of identical sequences (Shen and Huang 1986), the conserved regions of rRNA may be more prone to undergo recombination than protein-coding genes. In the latter, the redundancy of the genetic code permits synonymous substitutions, allowing for rapid divergence on the DNA sequence level that prevents homologous recombination (Gogarten et al. 2002). These recombination events are a feature that the rRNA may share with the genome as a whole: both are mosaics in which the different parts can have different histories.

2.2.3 Other Molecular Markers and Phylogentic Reconstruction Artifacts

Regardless of the difficulties, ribosomal RNA has become the gold standard for microbial taxonomy. As more molecular data became available, different gene trees were reconstructed and compared to rRNA trees. Some were in agreement with rRNA trees; others were not. One explanation for the incongruence of phylogenetic trees constructed using different markers is HGT; however, another important consideration is artifacts of phylogenetic reconstruction. The study of deep branching eukaryotic lineages reveals that one should not lose sight of the fact that the small subunit (SSU) rRNA-based tree of life at best depicts the evolutionary history of a single molecule. In some lineages this molecule evolved faster than in others. Many phylogenetic reconstruction algorithms tend to group long branches together, even if they are not specifically related (Felsenstein 1978). For example, the microsporidia, considered a deep-branching lineage based on SSU rRNA (Vossbrinck et al. 1987), have been recognized as a more recently emerging, rapidly evolving relatives of the fungi (Embley and Hirt 1998).

Branches in a molecular phylogeny are scaled with their lengths proportional to the amount of sequence change that occurred along a branch. Usually these branches are not scaled with respect to time. Based on the analyses of protein coding genes, the tree of eukaryotic evolution appears more bush-like than is the case in rRNA-based phylogenies. The clear distinction between crown group and deep branching eukaryotes appears to have been a particularity of ribosomal RNA (Simpson and Roger 2004).

3 Detecting HGT and Measuring Its Extent

Several types of phylogenetic and non-phylogenetic methods for detection of HGT events are being developed and constantly improved. Due to varying underlying assumptions, different methods detect HGTs at different phylogenetic distances and of different age, and therefore often return non-overlapping sets of HGT candidates (Ragan 2001a; Lawrence and Ochman 2002; Ragan 2002). In addition, all methods are imperfect and suffer from high rates of false positives and false negatives (Cortez et al. 2005).

3.1 Methods to Detect HGT

3.1.1 Approach #1: Surrogate Methods

Since a horizontally transferred gene comes from a different genomic background, its nucleotide sequence can contain signatures of its previous “home” genome. One group

of HGT detection methods uses either atypical nucleotide composition (Lawrence and Ochman 1997) and/or atypical codon usage patterns (Lawrence and Ochman 1998) to infer which genes in a genome are instances of HGT. Because these methods do not rely on phylogenetic reconstruction, they are sometimes called surrogate methods (Ragan 2001b). Because genes "ameliorate" (i.e., adapt to the signatures of its new genome) quickly (Lawrence and Ochman 1997), these methods are applicable to detection of very recent transfers only. While easily applicable to completely sequence genomes, these methods were criticized for returning high rates of false positives and negatives (Koski et al. 2001; Azad and Lawrence 2005; Cortez et al. 2005). An application of a compositional approach to 116 available genomes revealed that the number of recently transferred genes ranges from 0.5% in pea aphid endocellular symbiont *Buchnera* sp. APS to 25.2% in anaerobic methane-producing archaeon *Methanosarcina acetivorans* C2A (Nakamura et al. 2004).

Other surrogate methods are applicable only to very closely related organisms. Extent of HGT among the closely related organisms can be judged through a comparison of gene content of their genomes. For example, three sequenced *Escherichia coli* genomes share only 39.2% genes, while each *E. coli* genome separately has a substantial proportion of genes absent from two other strains (585 genes in non-pathogenic *E. coli* K12, 1,623 genes in uropathogenic *E. coli* CFT073 and 1,346 genes in enterohaemorrhagic *E. coli* O157:H7) (Welch et al. 2002).

3.1.2 Approach #2: Unusual Phyletic Patterns

Another way to assess whether a gene could have been transferred is to do a BLAST (Altschul et al. 1997) (or any other similarity or clustering algorithm) search of a sequence database (such as NCBI's non-redundant database) to find homologs to the query gene, define a gene family using this information and to look at the taxonomic distribution of members of the gene family (so called phyletic patterns). A significant top-scoring BLAST hit itself may suggest the most similar sequence in the database; this has been used to get a rough estimate of number of horizontally transferred genes in a genomes [e.g., bacterium *Thermotoga maritima* genome was proposed to have 24% of horizontally transferred genes from Archaea based on the top-scoring BLAST hits (Nelson et al. 1999)]. However, a top-scoring BLAST hit might not represent a sequence grouping with the query sequence, if a phylogenetic tree is reconstructed (Koski and Golding 2001), and therefore this is not a reliable approach for HGT detection. Phyletic patterns, however, can be further used to infer whether the patchy distribution of a gene is most parsimoniously explained by HGTs or by gains and losses (Snel et al. 2002; Kunin and Ouzounis 2003; Mirkin et al. 2003). The outcome of the inferences depends on a value of "HGT penalty" (a ratio between HGT events and gene losses), which is not known, but has to be estimated or set a priori, and different studies disagree on what value to use. A recent attempt to apply this type of approach to 165 microbial genomes resulted in an inference of ~40,000 horizontal gene transfers, ~90,000 gene losses and over ~600,000 vertical transfers in all analyzed gene families (Kunin and Ouzounis 2003). While the numbers given here may be interpreted as showing only a limited number of HGTs among the 165 genomes, one should not forget that those estimates do not consider HGTs resulting in orthologous replacement, which could constitute a substantial part of a genome (see approach #3).

3.1.3 Approach #3: Phylogenetic Incongruence

These methods rely on reconstruction of phylogenetic trees for sets of orthologous genes and comparison of them to each other, assuming that trees with unexpected (i.e., topologically incongruent) phylogenetic histories are results of horizontal gene transfer. These methods

typically use the expected phylogenetic history (organismal tree) as a reference tree for the comparison. One of the earliest such analyses came from comparison of gene families from *Aquifex aeolicus* genome to rRNA tree, with the conclusion that one gets "different phylogenetic placements based on what genes are used" (Pennisi 1998). Later, 205 gene families from 13 gamma-proteobacteria were compared to their concatenated phylogeny (Lerat et al. 2003) and 22,432 gene families from 144 prokaryotic genomes were compared to the supertree constructed from compatible bipartitions (Beiko et al. 2005). Choices of reference trees (as a proxy of organismal trees) include rRNA trees, genome trees, trees derived from concatenation of selected datasets, or trees (possibly only partially resolved) supported by a plurality of sets of orthologous genes (consensus trees or supertrees). Ideally, if the organismal tree is not known, all possible tree topologies should be tried as a reference tree [examples of such methodologies to analyze four and five genomes are in (Zhaxybayeva and Gogarten 2002; Zhaxybayeva et al. 2004a))]. However, this approach is computationally impossible for large-scale analyses, due the vast number of possible tree topologies. As an alternative, the trees to be analyzed could be broken into smaller pieces (e.g., bipartitions or quartets). There is a significantly smaller number of possible bipartitions/quartets than trees for a given number of analyzed genomes, and therefore all possibilities can be evaluated (giving rise to bipartition and quartet decomposition analyses) (Zhaxybayeva et al. 2004b, 2006). For example, analysis of 1,128 gene families in 10 cyanobacterial genomes resulted in 685 gene families with phylogenetic trees incongruent with a reference tree supported by a plurality of gene families (Zhaxybayeva et al. 2006), and hence providing candidates for HGTs.

One drawback of phylogenetic approaches (aside from artifacts of phylogenetic reconstruction, which are not discussed here) is that HGTs between neighboring taxa on the reference tree are invisible for these methods, because these transfers do not result in a change of tree topology. Another drawback is that a weak phylogenetic signal in a set of orthologous genes often results in an unresolved (or unsupported) tree topology. The latter topologies cannot be used to delineate HGT events, but they also should not be used [although unfortunately they are sometimes used, e.g. Snel et al. (2002)] as evidence for absence of HGT. The third drawback is that often a choice of reference tree may bias the results of HGT quantification. This is particularly a problem when a reference tree is obtained as a plurality tree (or supertree) from the same sets of genes that are subject to HGT detection in the study. The underlying assumption is that the number of HGTs should be minimized, and that the plurality of genes therefore reflects organismal evolution and not a reoccurring pattern of HGT.

3.2 Biological Consequences of HGT

3.2.1 Types of Transferred Genes

Jain et al. (1999) proposed that informational genes should be less likely to be transferred in comparison to operational genes because the former are part of highly interactive molecular assemblies. Although genome-wide analyses indicate that all types of genes are among inferred HGTs, certain functional categories are over- or under-represented among horizontally transferred genes as compared to their genome-wide distribution. In HGTs detected by a compositional method, Nakamura et al. (2004) reported bias of HGTs in functional categories of cell surface, DNA binding and pathogenicity related genes. In HGTs detected by a phylogenetic method, Beiko et al. (2005) reported over-representation of "energy metabolism" and "mobile and extrachromosomal element functions" among genes with discordant bipartitions in their phylogenies (i.e., among HGT candidates). However, transfer

distance may need to be taken into account when types of transferred genes are evaluated. For example, in genome-wide analyses of HGT in cyanobacteria, no bias toward a particular functional category is found for HGTs inferred to occur within cyanobacteria, but an excess of metabolic genes and decrease of informational genes is observed in transfers inferred to occur between cyanobacteria and other phyla (Zhaxybayeva et al. 2006).

3.2.2 How Tangled is the Web of Life?

Several recent large-scale genome analyses attempted to estimate HGT across all available sequenced genomes. Beiko et al. (2005) analyzed gene families from 144 prokaryotic genomes. They compared 95,194 strongly supported bipartitions from 22,432 gene family trees and found that 86.6% (82,473) of bipartitions are in agreement with bipartitions of a reference supertree (constructed from all highly supported bipartitions). Ge et al. (2005) analyzed 297 gene families from gene clusters of 40 microbial genomes and compared gene family phylogenies with a genome tree, applying an additional criteria to increase stringency. This resulted in 33 HGT events in 11.1% of 297 analyzed gene families. In the latter study, the investigators limited themselves to analyses of very strictly defined (ubiquitous) core genes, hence severely underestimating the number of HGT events. Analyzing larger number of gene families shows that in cyanobacteria as many as 61% of gene families may be affected by HGT (Zhaxybayeva et al. 2006). It should also be noted that all these analyses are based on phylogenetic approach (see Sect. 3.1.3) and hence ignore the pool of transitory genes (see Sect. 1.2.4), therefore underestimating the amount of gene transfer.

Nevertheless, attempts are still being made to resolve the tree of life using the vast amount of available genomic information. Recently Ciccarelli et al. (2006) performed an analysis of 191 genomes where they aimed to find genes with "indisputable orthology" and ended up with only 31 such genes (mainly ribosomal proteins). The joint analysis of these genes results in a phylogeny that has some resolution; however, the tree's backbone is poorly resolved ($<$80%), perhaps due to insufficient phylogenetic information and hence does not provide a clearly resolved strictly bifurcating tree. Dagan and Martin (2006) referred to this tree as "the tree of 1 percent" since only 1% of the genes were determined to fit the null hypothesis of a single strictly bifurcating tree of life. It is not sufficient to describe evolutionary relationships among only 1% of the genes and to claim that it represents an evolutionary history of the organisms.

3.2.3 Horizontal Gene Transfer and Genetic Life Rafts

As one moves down the tree of life, the branches tend to become sparse. Rather than suggesting that this reveals early life as less diverse, this might be a natural consequence of the large amount of extinction that has occurred over the history of life on Earth. Random models of extinction and speciation show that such sparse "deep branches" in phylogenies are expected, even when actual diversity remains constant over time (see Sect. 1.3). This observation has profound consequences in light of the large amount of horizontal gene transfer seen across today's extant lineages.

There is every reason to believe that HGT was at least as prevalent in the distant past as it is today (Woese 1998). However, as we move backward in time, it becomes less likely that the participants in any transfer event have any surviving descendants. Consequently, the more ancient the transfer event, the less likely that the donor lineage will still be in existence, even if the recipient has living descendants.

Furthermore, the more ancient the transfer event, the more divergent an extinct donor is likely to have been from currently existing lineages. In many cases, these genes may be the

only surviving biological signature from entire clades of extinct organisms, a "genetic life-raft". If these genes have any detectable homologs in existing species, the transferred gene will appear more deeply branching in a phylogeny of the gene family, even if the organismal tree shows the recipient lineage as highly derived. For example, the pyrrolysine aminoacyl-tRNA synthetase gene (PylS) is found only within a single family of derived Euryarchaea (the Methanosarcinales) and in one bacterial species (*Desulfitobacterium hafniense*) (Srinivasan et al. 2002). However, a phylogenetic tree of related aminoacyl-tRNA synthetases roots PylS more deeply than the MRCA of Bacteria and Archaea. This gene may well represent an ancient transfer from an extinct lineage. Since the amino acid pyrrolysine is only used at a single position within a specific set of enzymes unique to methanogenesis (the metabolic process of reducing single-carbon compounds to methane for energy production) that have no identified homologs (mono-, di-, and trimethylamine methyltransferases) (Ibba and Soll 2002), this may also suggest that this system must have originated outside the lineages that lead to extant methanogens. Evolution of a system to use an entirely novel amino acid would almost certainly require considerable positive selection, requiring (and resulting in) far more uses for Pyl. However, transfer of even a single advantageous gene requiring Pyl, along with the Pyl incorporation machinery, would result in its retention in the recipient lineage.

Perhaps even more interestingly, if there are no surviving homologs at all (as is the case with the methylamine methyltransferases), these genes will be listed as "orphans", and often may be uncharacterized simply for lack of comparative evidence from other more well-studied genes. Several pathways in methanogenesis contain many such orphans (Fournier, unpublished). As these pathways are found exclusively within a derived group of the Euryarchaea, the life-raft model provides a succinct explanation of their presence, which would otherwise have to be explained by either very rapid evolution, or ancestral presence with selective loss in other lineages.

Geological signatures and biochemical fossils (e.g., Schidlowski et al. 1983; Brocks et al. 2003) can be correlated to metabolic processes and physiological properties. However, in mapping these traits on to the tree of life (House et al. 2003), gene transfer, and in particular gene transfer form extinct lineages needs to be taken into consideration. A metabolic pathway might have existed a long time before it was transferred into an extant lineage; the geological signature or chemical fossil should not considered automatic proof that the lineage that today carries this trait was also responsible for creating the signature. Careful consideration of molecular phylogenies, including ancient gene duplications and gene transfers events, promises to yield a better understanding of the order in which metabolic pathways emerged.

4 Concluding Remarks

Advances in genome sequencing have revolutionized our understanding of microbial evolution. Reconstructing the evolutionary history of organisms turned out to be more difficult than anticipated two decades ago. Gene transfer and unequal substitution rates create artifacts that initially were not recognized.

Extending population genetic principles to larger taxonomic units helps to unravel evolutionary histories not only of the organisms, but also of the key pathways that allowed early organisms to survive on early Earth. The emerging picture of life's history less resembles a tree, but a coral formed by a tangle of lineages with many interconnections generated through the transfer of genes, or through the fusion of independent lines of descent. Most

of the organisms that form the bulk of this "coral of life" do not have extant representatives. However, some these extinct lineages transferred genes to lineages still alive today.

Acknowledgements Work in JPG's lab was supported the NSF (MCB-0237197), the NASA Applied Information Systems Research (NNG04GP90G) and NASA Exobiology Programs (NNX07AK15G). OZ is supported through a CIHR Postdoctoral Fellowship and is an honorary Killam Postdoctoral Fellow at Dalhousie University.

References

S.G. Acinas, L.A. Marcelino, V. Klepac-Ceraj, M.F. Polz, J. Bacteriol. **186**(9), 2629–2635 (2004)
S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. **25**(17), 3389–3402 (1997)
R.I. Amann, W. Ludwig, K.H. Schleifer, Microbiol. Rev. **59**(1), 143–169 (1995)
V. Anantharaman, E.V. Koonin, L. Aravind, Nucleic Acids Res. **30**(7), 1427–1464 (2002)
M. Arnold, *Evolution through Genetic Exchange* (Oxford University Press, Oxford, 2006)
R.K. Azad, J.G. Lawrence, PLoS Comput. Biol. **1**(6), e56 (2005)
R.G. Beiko, T.J. Harlow, M.A. Ragan, Proc. Natl. Acad. Sci. USA **102**(40), 14332–14337 (2005)
Y. Boucher, C.J. Douady, A.K. Sharma, M. Kamekura, W.F. Doolittle, J. Bacteriol. **186**(12), 3980–3990 (2004)
J.J. Brocks, R. Buick, R.E. Summons, G.A. Logan, Geochimica Cosmochimica Acta **67**(22), 4321–4335 (2003)
J.R. Brown, W.F. Doolittle, Proc. Natl. Acad. Sci. USA **92**(7), 2441–2445 (1995)
R.L. Cann, M. Stoneking, A.C. Wilson, Nature **325**(6099), 31–36 (1987)
F.D. Ciccarelli, T. Doerks, C. von Mering, C.J. Creevey, B. Snel, P. Bork, Science **311**(5765), 1283–1287 (2006)
J.R. Cole, B. Chai, R.J. Farris, Q. Wang, S.A. Kulam, D.M. McGarrell, G.M. Garrity, J.M. Tiedje, Nucl. Acids Res. **33**(Suppl. 1), D294–D296 (2005)
D.Q. Cortez, A. Lazcano, A. Becerra, In Silico Biol. **5**(5–6), 581–592 (2005)
T. Dagan, W. Martin, Genome Biol. **7**(10), 118 (2006)
C. Darwin, *Charles Darwin's Notebooks, 1836–1844, Transcription Published in 1987* (Cornell University Press, Ithaca, 1836–1844)
C. Darwin, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (John Murray, London, 1859)
V. Daubin, E. Lerat, G. Perriere, Genome Biol. **4**(9), R57 (2003)
R. Dawkins, *The Selfish Gene* (Oxford University Press, Oxford, 1976)
L. Delaye, A. Becerra, A. Lazcano, Orig. Life Evol. Biosphere **35**(6), 537–554 (2005)
P.P. Dennis, S. Ziesche, S. Mylvaganam, J. Bacteriol. **180**(18), 4804–4813 (1998)
H. Donoghue, M. Spigelman, Proc. R. Soc. B **273**(1587), 641–642 (2006)
T.M. Embley, R.P. Hirt, Curr. Opin. Genet. Dev. **8**(6), 624–629 (1998)
J. Felsenstein, Syst. Zool. **27**, 401–410 (1978)
J. Felsenstein, *Inferring Phylogenies* (Sinauer, Sunderland, 2003)
W.M. Fitch, K. Upper, Cold Spring Harb. Symp. Quant. Biol. **52**, 759–767 (1987)
G.M. Garrity, J.A. Bell, T.G. Lilburn, *Bergey's Taxonomic Outline* (Springer, New York, 2004). http://dx.doi.org/10.1007/bergeysoutline200310
F. Ge, L.-S. Wang, J. Kim, PLoS Biol. **3**(10), e316 (2005)
D. Gevers, F.M. Cohan, J.G. Lawrence, B.G. Spratt, T. Coenye, E.J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F.L. Thompson, J. Swings, Nat. Rev. Microbiol. **3**(9), 733–739 (2005)
A. Gibbons, Science **310**(5756), 1889 (2005)
J.P. Gogarten, W.F. Doolittle, J.G. Lawrence, Mol. Biol. Evol. **19**(12), 2226–2238 (2002)
J.P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E.J. Bowman, B.J. Bowman, M.F. Manolson, R.J. Poole, T. Date, T. Oshima et al., Proc. Natl. Acad. Sci. USA **86**(17), 6661–6665 (1989)
J.P. Gogarten, L. Taiz, Photosynth. Res. **33**, 137–146 (1992)
J.P. Gogarten, J.P. Townsend, Nat. Rev. Microbiol. **3**(9), 679–687 (2005)
M. Gogarten-Boekels, E. Hilario, J.P. Gogarten, Orig. Life Evol. Biosphere **25**(1–3), 251–264 (1995)
R. Gomes, H.F. Levison, K. Tsiganis, A. Morbidelli, Nature **435**(7041), 466–469 (2005)
P.R. Grant, B.R. Grant, J.A. Markert, L.F. Keller, K. Petren, Evolution **58**(7), 1588–1599 (2004)
S. Gribaldo, P. Cammarano, J. Mol. Evol. **47**(5), 508–516 (1998)

J. Gurevitch, D.K. Padilla, Trends Ecol. Evol. **19**(9), 470–474 (2004)
E. Haeckel, *Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft mechanisch begründet durch die von Charles Darvin reformierte Descendenz-Theorie* (Georg Riemer, Berlin, 1866)
W.P. Hanage, C. Fraser, B.G. Spratt, BMC Biol. **3**(1), 6 (2005)
C.H. House, B. Runnegar, S.T. Fitz-Gibbon, Geobiology **1**(1), 15–26 (2003)
W.W. Hsiao, K. Ung, D. Aeschliman, J. Bryan, B.B. Finlay, F.S. Brinkman, PLoS Genet. **1**(5), e62 (2005)
J. Huang, Y. Xu, J.P. Gogarten, Mol. Biol. Evol. **22**(11), 2142–2146 (2005)
M. Ibba, D. Soll, Curr. Biol. **12**(13), R464–R466 (2002)
N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, T. Miyata, Proc. Natl. Acad. Sci. USA **86**(23), 9355–9359 (1989)
R. Jain, M.C. Rivera, J.A. Lake, Proc. Natl. Acad. Sci. USA **96**(7), 3801–3806 (1999)
R. Jain, M.C. Rivera, J.E. Moore, J.A. Lake, Mol. Biol. Evol. **20**(10), 1598–1602 (2003)
K.T. Konstantinidis, J.M. Tiedje, Proc. Natl. Acad. Sci. USA **102**(7), 2567–2572 (2005)
L.B. Koski, G.B. Golding, J. Mol. Evol. **52**(6), 540–542 (2001)
L.B. Koski, R.A. Morton, G.B. Golding, Mol. Biol. Evol. **18**(3), 404–412 (2001)
V. Kunin, C.A. Ouzounis, Bioinformatics **19**(11), 1412–1416 (2003)
J.-B. Lamarck, *Histoire naturelle des animaux sans vertebras* (Paris, 1815)
B.F. Lang, C. O'Kelly, T. Nerad, M.W. Gray, G. Burger, Curr. Biol. **12**(20), 1773–1778 (2002)
P. Lapierre, R. Shial, J.P. Gogarten, Syst. Appl. Microbiol. **29**(1), 15–23 (2006)
J.G. Lawrence, H. Hendrickson, Mol. Microbiol. **50**(3), 739–749 (2003)
J.G. Lawrence, H. Ochman, J. Mol. Evol. **44**(4), 383–397 (1997)
J.G. Lawrence, H. Ochman, Proc. Natl. Acad. Sci. USA **95**(16), 9413–9417 (1998)
J.G. Lawrence, H. Ochman, Trends Microbiol. **10**(1), 1–4 (2002)
E. Lerat, V. Daubin, N.A. Moran, PLoS Biol. **1**(1), E19 (2003)
R.E. Ley, J.K. Harris, J. Wilcox, J.R. Spear, S.R. Miller, B.M. Bebout, J.A. Maresca, D.A. Bryant, M.L. Sogin, N.R. Pace, Appl. Environ. Microbiol. **72**(5), 3685–3695 (2006)
M. Lynch, Annu. Rev. Microbiol. **60**, 327–349 (2006)
M. Lynch, J.S. Conery, Science **302**(5649), 1401–1404 (2003)
L. Margulis, *Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons* (Freeman, 1995)
A.P. Martin, E.K. Costello, A.F. Meyer, D.R. Nemergut, S.K. Schmidt, Evol. Int. J. Org. Evol. **58**(5), 946–955 (2004)
W. Martin, E.V. Koonin, Nature **440**(7080), 41–45 (2006)
E. Mayr, *Systematics and the Origin of Species* (Columbia Univ. Press, New York, 1942)
B.G. Mirkin, T.I. Fenner, M.Y. Galperin, E.V. Koonin, BMC Evol. Biol. **3**(1), 2 (2003)
A. Morandi, O. Zhaxybayeva, J.P. Gogarten, J. Graf, J. Bacteriol. **187**(18), 6561–6564 (2005)
S. Mylvaganam, P.P. Dennis, Genetics **130**(3), 399–410 (1992)
Y. Nakamura, T. Itoh, H. Matsuda, T. Gojobori, Nat. Genet. **36**(7), 760–766 (2004)
K.E. Nelson, R.A. Clayton, S.R. Gill, M.L. Gwinn, R.J. Dodson, D.H. Haft, E.K. Hickey, J.D. Peterson, W.C. Nelson, K.A. Ketchum, L. McDonald, T.R. Utterback, J.A. Malek, K.D. Linher, M.M. Garrett, A.M. Stewart, M.D. Cotton, M.S. Pratt, C.A. Phillips, D. Richardson, J. Heidelberg, G.G. Sutton, R.D. Fleischmann, J.A. Eisen, O. White, S.L. Salzberg, H.O. Smith, J.C. Venter, C.M. Fraser, Nature **399**(6734), 323–329 (1999)
E.G. Nisbet, N.H. Sleep, Nature **409**(6823), 1083–1091 (2001)
L.O. Olendzenski, L. Liu, O. Zhaxybayeva, R. Murphey, D.G. Shin, J.P. Gogarten, J. Mol. Evol. **51**(6), 587–599 (2000)
L.O. Olendzenski, O. Zhaxybayeva, J.P. Gogarten, in *Microbial Genomes*, ed. by C.M. Fraser, T. Read, K. Nelson (Humana Press, 2004), pp. 143–154
R.D.M. Page, E.C. Holmes, *Molecular Evolution: A Phylogenetic Approach* (Blackwell, 1998)
E. Pennisi, Science **280**(5364), 672–674 (1998)
J. Pereto, P. Lopez-Garcia, D. Moreira, Trends Biochem. Sci. **29**(9), 469–477 (2004)
H. Philippe, E.A. Snell, E. Bapteste, P. Lopez, P.W. Holland, D. Casane, Mol. Biol. Evol. **21**(9), 1740–1752 (2004)
M.A. Ragan, Curr. Opin. Genet. Dev. **11**(6), 620–626 (2001a)
M.A. Ragan, FEMS Microbiol. Lett. **201**(2), 187–191 (2001b)
M.A. Ragan, Trends in Microbiol. **10**(Suppl. 1), 4 (2002)
D.M. Raup, Paleobiology **11**(1), 42–52 (1985)
D.M. Raup, Philos. Trans. R. Soc. Lond. B. Biol. Sci. **325**, 421–431 (1989); discussion 431–435
D.L. Rohde, S. Olson, J.T. Chang, Nature **431**(7008), 562–566 (2004)
A. Rokas, S.B. Carroll, PLoS Biol. **4**(11), e352 (2006)

R. Rossello-Mora, R. Amann, FEMS Microbiol. Rev. **25**(1), 39–67 (2001)
J. Sapp, Microbiol. Mol. Biol. Rev. **69**(2), 292–305 (2005)
M. Schidlowski, J.M. Hayes, I.R. Kaplan, in *Earth's Earliest Biosphere*, ed. by J.W. Schopf (Princeton University Press, Princeton, 1983), pp. 149–187.
P.D. Schloss, J. Handelsman, Microbiol. Mol. Biol. Rev. **68**(4), 686–691 (2004)
P. Shen, H.V. Huang, Genetics **112**(3), 441–457 (1986)
A.G. Simpson, A.J. Roger, Curr. Biol. **14**(17), R693–R696 (2004)
B. Snel, P. Bork, M.A. Huynen, Genome Res. **12**(1), 17–25 (2002)
M.L. Sogin, H.G. Morrison, J.A. Huber, D.M. Welch, S.M. Huse, P.R. Neal, J.M. Arrieta, G.J. Herndl, Proc. Natl. Acad. Sci. USA **103**(32), 12115–12120 (2006)
G. Srinivasan, C.M. James, J.A. Krzycki, Science **296**(5572), 1459–1462 (2002)
R.Y. Stanier, C.B. Van Niel, Arch. Mikrobiol. **42**, 17–35 (1962)
R. Thomson, J.K. Pritchard, P. Shen, P.J. Oefner, M.W. Feldman, Proc. Natl. Acad. Sci. USA **97**(13), 7360–7365 (2000)
P.A. Underhill, P. Shen, A.A. Lin, L. Jin, G. Passarino, W.H. Yang, E. Kauffman, B. Bonne-Tamir, J. Bertranpetit, P. Francalacci, M. Ibrahim, T. Jenkins, J.R. Kidd, S.Q. Mehdi, M.T. Seielstad, R.S. Wells, A. Piazza, R.W. Davis, M.W. Feldman, L.L. Cavalli-Sforza, P.J. Oefner, Nat. Genet. **26**(3), 358–361 (2000)
L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, A.C. Wilson, Science **253**(5027), 1503–1507 (1991)
C.R. Vossbrinck, J.V. Maddox, S. Friedman, B.A. Debrunner-Vossbrinck, C.R. Woese, Nature **326**(6111), 411–414 (1987)
R.A. Welch, V. Burland, G. Plunkett, 3rd, P. Redford, P. Roesch, D. Rasko, E.L. Buckles, S.R. Liou, A. Boutin, J. Hackett, D. Stroud, G.F. Mayhew, D.J. Rose, S. Zhou, D.C. Schwartz, N.T. Perna, H.L. Mobley, M.S. Donnenberg, F.R. Blattner, Proc. Natl. Acad. Sci. USA **99**(26), 17020–17024 (2002)
E. Willerslev, A. Cooper, Proc. R. Soc. B: Biol. Sci. **272**(1558), 3–16 (2005)
C. Woese, Proc. Natl. Acad. Sci. USA **95**(12), 6854–6859 (1998)
C. Woese, O. Kandler, M. Wheelis, Proc. Natl. Acad. Sci. USA **87**, 4576–4579 (1990)
C.R. Woese, Microbiol. Rev. **51**(2), 221–271 (1987)
C.R. Woese, G.E. Fox, Proc. Natl. Acad. Sci. USA **74**(11), 5088–5090 (1977)
W.H. Yap, Z. Zhang, Y. Wang, J. Bacteriol. **181**(17), 5201–5209 (1999)
O. Zhaxybayeva, J. Gogarten, BMC Genomics **3**, 4 (2002)
O. Zhaxybayeva, J.P. Gogarten, Trends Genet. **20**(4), 182–187 (2004)
O. Zhaxybayeva, J.P. Gogarten, R.L. Charlebois, W.F. Doolittle, R.T. Papke, Genome Res. **16**(9), 1099–1108 (2006)
O. Zhaxybayeva, L. Hamel, J. Raymond, J. Gogarten, Genome Biol. **5**(3), R20 (2004a)
O. Zhaxybayeva, P. Lapierre, J.P. Gogarten, Trends Genet. **20**(5), 254–260 (2004b)
O. Zhaxybayeva, P. Lapierre, J.P. Gogarten, Protoplasma **227**(1), 53–64 (2005)