Lawrence Berkeley National Laboratory

Recent Work

Title

Phylogenetic relationships of bacteria with special reference to Endosymbionts and Enteric Species

Permalink

<https://escholarship.org/uc/item/2nr448pv>

Authors

Francino, M. Pilar
Santos, Scott R.
Ochman, Howard

Publication Date

1992

LBNL-52742

**PHYLOGENETIC RELATIONSHIPS OF BACTERIA WITH
SPECIAL REFERENCE TO
ENDOSYMBIONTS AND ENTERIC SPECIES**

M. Pilar Francino¹, Scott R. Santos² and Howard Ochman²

¹Laboratory of Genomic Diversity, DOE Joint Genome Institute, Walnut Creek, CA

94598 ²Department of Biochemistry & Molecular Biophysics, University of Arizona,
Tucson, AZ 85721

For: M. Dworkin (Editor-in-Chief) *The Prokaryotes 3rd edition, a Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications*. Springer-Verlag, NY

**PHYLOGENETIC RELATIONSHIPS OF BACTERIA WITH SPECIAL REFERENCE TO
ENDOSYMBIONTS AND ENTERIC SPECIES**

M. Pilar Francino¹, Scott R. Santos² and Howard Ochman²

¹Laboratory of Genomic Diversity, DOE Joint Genome Institute, Walnut Creek, CA 94598

²Department of Biochemistry & Molecular Biophysics, University of Arizona, Tucson, AZ

85721

Address for correspondence:

Howard Ochman, Department of Biochemistry and Molecular Biophysics,
233 Life Sciences South, University of Arizona, Tucson, Arizona 85721, USA
Phone: 520-626-8355, Fax: 520-621-3709, E-mail: hochman@email.arizona.edu

INTRODUCTION

The latter half of the 20th century saw two developments that revolutionized the study of bacterial systematics and our conceptions about microbial diversity. The first was the use of molecular genetic information, as gained through the analysis of protein and nucleic acid sequences, for the identification, typing and classification of microorganisms. The second was the application of PCR for the recovery of DNA sequences from non-cultivable organisms.

By decoupling cultivation and classification, and by providing a common yardstick by which all organisms could be compared (*i.e.*, small subunit ribosomal DNA sequences; hereafter referred to as SSU rRNA), these procedures added a new dimension to our knowledge of the microbial world. It is no longer necessary to test organisms against a panel of known phenotypic or biochemical properties, which, in themselves, are often variable in occurrence and can lead to cases of misclassification, or carry a preconceived notion about the characters that define a particular taxonomic group.

Despite the considerable advances made from applying these innovations, such an approach is not without its limitations. Characterization of a single conserved sequence, while often adequate for taxonomic purposes, does not reveal very much about the evolutionary history or biology of the organism *per se*. Whereas the placement of an organism (*i.e.*, sequence) into an evolutionary framework may prompt numerous predictions about its morphology, ecology, physiology and biochemistry, virtually nothing can be said about its unique biology without cultivation or additional sequence information.

Other problems can arise when attempting to classify bacteria and to determine their relationships. These derive from two sources: the specific characters used to represent an

organism (the choice of data), and the algorithms used to infer relationships (the methods used to construct the phylogeny or tree). With this in mind, genotypic data hold several advantages over phenotypic characters for phylogenetic reconstruction in that: (i) variation in molecular sequences is discrete and well-defined; (ii) molecular sequences comprise hundreds, if not thousands, of characters bearing potentially useful information; (iii) molecular sequences accumulate informative changes, even when the resulting phenotypes are conserved; (iv) informational macromolecules (DNA and proteins) change according to well-defined models of sequence evolution; and (v) identical phenotypes can arise by very different genetic mechanisms or pathways, such that changes other than those attributable to common ancestry can lead to the same phenotype.

Questions of character convergence are not limited to phenotypic traits: portions of molecular sequences may be identical due to chance, natural selection, or lateral gene transfer. Furthermore, because the analysis of relatively small regions of DNA can furnish so much information, phylogenetic relationships are often based on data collected from a very limited portion of the genome, which may or may not be representative of the organism as a whole. Such problems stress the need to examine multiple genes and to infer statistically robust phylogenetic trees, which assists in recognizing inconsistencies in the data and in establishing an evolutionary framework for studying the biology of an organism.

A GENOMIC VIEW OF BACTERIAL EVOLUTION

The first goal of this chapter is to examine how information coming from completely sequenced microbial genomes has contributed to our interpretations of bacterial phylogeny and evolution. The molecular evolutionary history of microbes has largely been based on the evaluation of SSU rRNA (Fox et al. 1977, 1980, Woese 1987, Hugenholtz et al. 1998, Maidak et

al. 2000); however, the analysis of several conserved protein-coding regions has offered alternative scenarios relating to the branching orders of bacterial taxa, the rooting of the universal tree of life, and the role of lateral transfer in the early evolution and formation of genomes (Gogarten et al. 1989, Iwabe et al. 1989, Hilario and Gogarten 1993, Brown and Doolittle 1995, Golding and Gupta 1995, Baldauf et al. 1996, Eisen 1998, Forterre and Phillipe 1999, Schutz et al. 2000). Such results have promoted the view that SSU rRNA might not fully reveal the relationships among organisms because genomes contain genes from multiple sources. Quite apart from the fact that SSU rRNA spans a very small portion and may not be representative of an entire genome, other concerns have been raised about the use of this molecule as the sole basis for molecular phylogenetic reconstructions. The size of SSU rRNA (ca.1600 nucleotides) and its very slow rate of evolution contribute to the fact that there are often too few variable sites to accurately resolve phylogenetic relationships. Furthermore, because rRNA does not encode a protein, the occurrence of frequent additions and deletions in certain portions of the molecule presents problems for sequence alignments, and selection on secondary structures can cause sequence convergence and saturation (homoplasy), thereby distorting the actual relationships among organisms (Hillis and Dixon 1991, Kjer 1995, Huynen et al. 1997).

Complete genomes allow immediate access to the full set of sequences that are conserved among diverse genomes, which has greatly facilitated the analysis of genes other than those encoding SSU rRNA. Perhaps a hundred genes are conserved among all bacterial lineages, affording the opportunity to trace the evolutionary history of sequences representing a substantial fraction of the genome. Furthermore, the availability of complete genome sequences has led to the development of numerous “whole-genome” approaches for evaluating the relationships among organisms. These studies are certainly among the most interesting endeavors in the

analysis of bacterial genomes and have broadened the scope of evolutionary studies to include new levels of genetic organization.

The second portion in this chapter concerns the phylogenetic relationships among microorganisms classified as Enterobacteriaceae. Considering that this family of bacteria includes the workhorses of molecular, genetic, and biochemical analyses (*E. coli* and *Salmonella*) and contains pathogens causing diseases with substantial impact on human populations (plague, typhoid, dysentery), there is perhaps more known about the biology of these microbes than for any other group of organisms. As in many other taxa, there has been a shift towards the use of molecular characters to infer the relationships among enterics; however, phenotypic markers still form much of the basis for typing the majority of strains, particularly in clinical settings. Because most enteric species were originally differentiated based on biotyping rather than on genotypic characters, several of the current designations are arbitrary and do not reflect either the extent of genetic diversity or the true relationships among strains.

As an example, the Shigellae have traditionally been subdivided into four species – *Shigella boydii*, *S. flexneri*, *S. dysenteriae* and *S. sonnei* – although the total amount of genetic variation within this genus is fully contained within *Escherichia coli*. Therefore, *Shigella* should not be classified as a genus separate from *Escherichia*. Furthermore, the amount of genetic diversity varies widely among the four *Shigella* species: *S. sonnei* constitutes a single genetically uniform lineage, whereas each of the other three species comprises a very heterogeneous array of clones. Finally, the classification of *Shigella* strains does not reflect their genetic similarities: strains of *S. flexneri* can be more closely related to *S. boydii* than to any other strains typed as *S. flexneri*, and both *Shigella boydii* and *S. dysenteriae* have multiple independent (polyphyletic) origins

from within *E. coli*. Clearly, the Shigellae warrant reclassification based on genetic criteria, although the epidemiological consequences of a taxonomic revision are uncertain.

Although there are numerous cases where traditional classification is at odds with information from molecular sequences, or where the phylogenetic relationships based on different molecular characters are in conflict, our purpose in this chapter is not to amend or revise the taxonomic nomenclature either of the Enterobacteriaceae or of the major phyla of Bacteria. Rather, we outline the rationale of the phylogenetic approach as it applies to bacteria, describe several of the molecular characters and methods used to infer phylogenetic trees, and review the phylogenetic relationships of strains derived from these procedures. Hopefully, in doing so, we have accomplished our ultimate goal, which is to provide an evolutionary framework for understanding the remarkable diversity of bacteria.

BACTERIAL RELATIONSHIPS AND REALITIES

The living world is organized taxonomically into three domains – Archaea, Bacteria, and Eucarya – which have been delineated in terms of their SSU rRNA sequences and cell membrane structures (Woese et al. 1990, Winker and Woese 1991). Despite the prevailing nomenclature, Bacteria are the most ancient and are estimated to have appeared between 3.5 and 4 billion years ago. Bacteria comprise about 40 phyla (*i.e.*, deep-branching subdivisions), about one-third of which presently contain representatives among the fully sequenced genomes. Naturally, cultivable, pathogenic, and small-genome bacteria have been the focus of most sequencing endeavors; however, substantial efforts are being made to expand both the phylogenetic and ecological scope of the set of sequenced microbes.

For assessing the broad-scale evolutionary relationships among the Bacteria, this chapter considers primarily those microbes for which full genome sequences are available. Although this limitation overlooks large portions of the genomic diversity within Bacteria, it is useful to note that: (i) many recognized bacterial phyla are represented by only one or a few organisms for which we have little more than SSU rRNA sequences; and (ii) the early radiation of Bacteria into phyla occurred relatively rapidly such that the set of sequenced organisms probably encompasses the range of variation within this domain.

The field of phylogenetics is constantly being challenged with new types of data, larger samples of organisms, and alternative models of evolution, and thus relies upon the continuous development of new methods to estimate, infer, test and resolve the relationships among organisms. Virtually all phylogenetic trees encountered in the literature are based on one of the following methods: (i) *Parsimony* finds the tree that requires the minimum number of changes; (ii) *Simple Distance* generates the tree that best represents the matrix of pairwise distances between organisms; (iii) *Neighbor Joining* is a distance method that progressively clusters pairs of taxa that yield the shortest total branch length; (iv) *Maximum likelihood* uncovers the tree that best fits a specified probabilistic model of sequence evolution; and (v) *Bayesian inference* relies upon the posterior probabilities of a tree to test the hypothesis that the particular topology is correct. Because parsimony and likelihood methods compare all possible trees in order to find the optimal branching order, both become computationally intensive with increasing sample sizes. *Bootstrapping* is a data re-sampling procedure designed to test the degree of support for the branching order in a given tree, with numbers (bootstrap values) placed at nodes to indicate the percentage of replicates in which the same taxa are joined at a particular node.

SEQUENCE-ALIGNMENT-BASED BACTERIAL PHYLOGENIES

Since the early 1980's, when small-scale DNA sequencing became affordable, bacterial phylogenetics has relied upon the comparison of homologous sequences in different taxa and, in particular, on SSU rRNA. Such comparisons have allowed for an ability to comprehend the diversity within the microbial world, and to define major phylogenetic groups. The preponderance of SSU rRNA as a phylogenetic character stemmed mostly from the fact that it is universally distributed and slowly evolving, and performs essentially the same function in all life forms. However, ribosomal RNA is not unique in this respect in that many proteins have a very wide distribution, with over 100 protein-coding genes common to all Bacteria. Analyses based on different subsets of these shared genes, ranging from single genes to extensive sequence concatenations, have resulted in trees whose topologies differ in many respects from each other and from the SSU rRNA tree. The current challenges are: (i) to recognize among the wealth of sequence information available the subset of genes, or genomic characters, that carry the least distorted phylogenetic signal and (ii) to develop methods that allow for the meaningful integration of information distilled from a wide number of genes.

The SSU rRNA bacterial phylogeny. The SSU rRNA tree for Bacteria currently delineates about 40 phyla, of which 13 are formed by uncultured species represented only by sequences directly cloned from the environment (Hugenholtz et al. 1998a, b). However, the relationships among these phyla are very poorly resolved, so that, except for a few deep branching lineages, the SSU rRNA phylogeny resembles a “bush” rather than a tree, which is often interpreted as evidence for a rapid adaptive radiation and diversification of early bacterial life (Woese 1987, Hugenholtz et al. 1998a, b, Cavalier-Smith 2002). Moreover, the addition of increasing numbers of taxa to the SSU rRNA tree has resulted in a failure to significantly support a single origin (*i.e.*,

monophyly) for some traditionally accepted phyla, such as the Proteobacteria. Indeed, although SSU rRNA clearly identifies a clade uniting the gamma-, beta- and alpha-Proteobacteria (with the first two as sister groups), it does not support the inclusion of the delta- and epsilon-Proteobacteria (Ludwig and Klenk 2001). The best-supported higher-level grouping in the SSU rRNA tree unites the phyla Chlorobi (green sulfur bacteria) and Bacteroidetes (Cytophaga/Flexibacter/Bacteroides group). Rooting of the tree with archaeal and eukaryotic sequences suggests that Aquificae is the most deeply branching bacterial phylum, followed by a number of phyla represented only by environmental sequences, the Thermodesulfobacteria, the Dictyoglomi and the Thermotogae (fig. 1). All of these deeply branching phyla are thermophilic, which, along with the fact that thermophyly is thought to be ancestral in Archaea, has been taken as evidence for a “hot” origin of life (Pace 1991). A reanalysis of SSU rRNA alignments restricted to slowly evolving positions, which are less prone to multiple substitutions and may be more reliable for early divergences (Brochier and Philippe 2002), groups Aquificae and Thermotogae in a derived position and places the phylum Planctomycetes as the basal lineage of Bacteria.

Contrasting SSU rRNA phylogeny with other phylogenetic markers. Does the topology of the SSU rRNA tree reflect a true evolutionary explosion of phyla or does this topology result from a lack of resolving power of this molecule at this deep level? These questions have been addressed by the analysis of additional genes and proteins that are shared by all bacteria. Bacterial single-gene phylogenies have been produced for a considerable number of molecules, including large subunit (LSU; 23S) rRNA (De Rijk et al. 1995, Ludwig et al. 1998, Ludwig and Klenk 2001), elongation factor EF-Tu (Ludwig et al. 1993, Delwiche et al. 1995, Baldauf et al. 1996, Gupta 1998, Ludwig et al. 1998, Bocchetta et al. 2000, Jenkins and Fuerst 2001, Ludwig and Klenk 2001), RNA polymerase (Klenk et al. 1999, Bocchetta et al. 2000) and sigma factor (Gruber and

Bryant 1997), RecA (Eisen 1995, Gruber et al 1998), cytochrome (Schutz et al. 2000) and ATPase subunits (Ludwig et al. 1993,1998, Brown and Doolittle 1997, Ludwig and Klenk 2001), and heat shock proteins HSP60 (Viale et al. 1994) and HSP70 (Brown and Doolittle 1997, Gupta et al. 1997 1998). In general, no other single phylogenetic marker provides resolution much beyond the SSU rRNA tree. Most single-gene phylogenies support the existence of the phyla delineated by SSU rRNA, and, in many cases, the basal placement of the hyperthermophilic phyla. Occasionally, single-gene phylogenies support the Chlorobi-Bacteroidetes relationship seen in the rRNA tree, or indicate novel clades, sometimes in contradiction with the rRNA phylogeny. The salient features of some of these trees are summarized below.

Other translational molecules. The phylogeny obtained with LSU rRNA is virtually identical to that of SSU (De Rijk et al. 1995, Ludwig et al. 1998, Ludwig and Klenk 2001), which might be expected given the functional coupling of these molecules and their potential for being affected by the same biases on sequence evolution (GC pressure, selection on secondary structure, etc). Although the LSU is longer and contains more potentially informative sites than the SSU, its resolution power at the level of inter-phyla relationships is equally poor. However, Brochier et al. (2002) have analysed a concatenated data set containing both rRNAs and obtained some interesting higher-level groupings. In this tree, which includes only species whose genomes have been sequenced, the epsilon-Proteobacteria are included along with the other proteobacterial classes with high bootstrap support (but the delta-Proteobacteria were not represented in the data set). Two clades strongly supported by the rRNA concatenation are that of Cyanobacteria and *Deinococcus* and that of the hyperthermophilic phyla Aquificae and Thermotogae, represented by the genome sequences of *Aquifex* and *Thermotoga*. In addition, in the concatenated rRNA tree, the hyperthermophiles are linked with the Actinobacteria. A close

affiliation between *Thermotoga* and Gram positive bacteria (Actinobacteria and Firmicutes) has been defended by Cavalier-Smith (1992, 2002) on ultrastructural grounds and has been suggested by phylogenetic analyses based on different proteins (Tiboni et al. 1993, Brown et al. 1994, Karlin 1995, Gupta 1998).

The translation elongation factor EF-Tu is, like rRNAs, a universally distributed component of the translation machinery, and highly conservative in sequence and function due to its basic role in the elongation of nascent polypeptides. Many phylogenetic reconstructions have employed this marker and, overall, they are in good agreement with the rRNA trees (Ludwig et al. 1993, Delwiche et al. 1995, Baldauf et al. 1996, Gupta 1998, Ludwig et al. 1998, Bocchetta et al. 2000, Jenkins and Fuerst 2001, Ludwig and Klenk 2001). Unfortunately, the resolving power of EF-Tu is inferior to that of rRNA on account of the limited numbers of informative sites, and although it supports the clustering of Chlorobi and Bacteroidetes, it does not provide any further insight into inter-phyla relationships. Even a phylogeny based on a concatenation of EF-Tu and its paralog EF-G fails to recover the monophyly of Proteobacteria or any higher-level grouping, other than a clade uniting all phyla with the exception of the deeply branching *Deinococcus-Thermus*, *Thermotogae* and *Aquificae* (Bocchetta et al. 2000).

Information-processing proteins. Other well-conserved, information-processing molecules have been used for phylogenetic reconstruction, including RNA polymerase (Klenk et al. 1999, Bocchetta et al. 2000) and its σ^{70} -type sigma factor (Gruber and Bryant 1997), and RecA (Eisen 1995, Gruber et al 1998). RNA polymerase (RNAP) presents the added advantage of being devoid of cases of paralogy, which often complicate protein-based phylogenies. However, the trees based on RNAP subunits are equivocal and differ from the rRNA tree in the position of the

hyperthermophiles and the mycoplasmas. Instead of their commonly recovered placement at the base of the bacterial tree, Aquificae and Thermotogae branch in the proximity of Spirochaetes and Proteobacteria, while the mycoplasmas appear as the most divergent bacterial lineage. However, this result is likely an artifact resulting from the skewed base-composition and high rate of evolution of mycoplasmas because Aquificae and Thermotogae return to their usual position at the base of the tree when the mycoplasmas are removed from analysis (Bocchetta 2000).

The phylogeny based on the σ^{70} -type sigma factor (a dissociable subunit of the RNA polymerase holoenzyme that confers promoter specificity) also displaces the Thermotogae from the base of the tree, grouping it with *Mycoplasma*, albeit with low support (Gruber and Bryant 1997). In addition, this phylogeny strongly groups a monophyletic Proteobacteria (although the epsilon class is not represented) with Chlamydiae.

RecA is a versatile protein, involved in homologous DNA recombination and SOS induction. The resolving power of RecA is higher than that of other proteins in that it recognizes the monophyly of Proteobacteria, although it can not resolve the relative branching order of the delta and epsilon classes. The RecA-based phylogeny corroborates very strongly the relationship between Chlorobi and Bacteroidetes, as seen in rRNA and EF-Tu trees, but it does not provide any further insight into inter-phyla relationships (Eisen 1995, Gruber et al 1998).

Proteins involved in metabolism and other cellular processes. Numerous phylogenies have been obtained for molecules in non-informational roles, including the bioenergetics-related ATPases and cytochromes, and chaperonins. In the case of the ATP-ase β -subunit (Ludwig et al. 1993,

1998, Brown and Doolittle 1997, Ludwig and Klenk 2001), the resolving power of this protein is so low that monophyly is not recovered for such phyla as the Proteobacteria and the Firmicutes. On the other hand, a phylogeny based on subunits of cytochrome bc complexes contradicts the rRNA tree by grouping the Aquificae with the Proteobacteria (Schutz et al. 2000), as is also seen in some RNAP trees (Klenk et al. 1999, Bocchetta et al. 2000). Although the position of the Aquificae in these trees may be due to homoplasy or horizontal transfer (Bocchetta et al. 2000, Schutz et al. 2000), other evidence hints to a relationship between the Aquificae and Proteobacteria. These two phyla share a long insertion in the RNAP β subunit not present in any other bacterial phylum (Klenk et al. 1999), as well as an insertion in alanyl-tRNA synthetase also seen in Chlamydiae, Chlorobi and Bacteroidetes (Gupta et al. 1999). These facts have led several authors to suggest that the Aquificae are in fact close relatives of Proteobacteria, and that their frequent position at the base of the bacterial tree is in reality an artifact caused by long-branch-attraction (Philippe and Laurent 1998, Gupta et al. 1999, Cavalier-Smith 2002).

The phylogenies based on the evolutionarily unrelated chaperonins HSP60 (or GroEL; Viale et al. 1994) and HSP70 (Brown and Doolittle 1997, Gupta et al. 1997, Gupta 1998) are specially interesting because they are the sole single-gene phylogenies that have been reported to divide relationships in the tight “bush”-like zone of the Bacterial tree. HSP60 partitions the bacterial phyla into two well-separated clusters: 1) the Proteobacteria, Chlamydiae, Spirochaetes and Bacteroidetes, and 2) the Cyanobacteria, Actinobacteria and Firmicutes. In contrast, analyses by Gupta and coworkers (Gupta et al. 1997, Gupta 1998) have found that chaperonin HSP70 includes the Cyanobacteria in cluster 1. In these phylogenies, the *Deinococcus-Thermus* phylum is strongly linked to Cyanobacteria, which results in a well-supported grouping of all gram-negative phyla, and the two gram-positive phyla, Actinobacteria and Firmicutes, are also united

with relatively high support. However, archaeal genera also cluster with the Gram-positives, rendering the reliability of these HSP70-based phylogenies dubious. Moreover, a similar HSP70 analysis by Brown and Doolittle (1997) does not substantiate any significant high-level groupings among bacterial phyla.

Post-genomic phylogenies based on alignments of extensive data sets. Clearly, no single molecule possesses sufficient resolving power to clarify the ancient relationships among bacterial phyla. Moreover, the conflicts between phylogenies obtained with different markers and the growing awareness of the contribution of horizontal gene transfer to the bacterial genome indicate that phylogenies based on single molecules may not be representative of the evolution of organisms, and have even cast doubts on the existence of an organismal bacterial phylogeny (Gogarten 1995, Doolittle 1999, Zhaxybayeva and Gogarten 2002). The recent availability of genomic sequences for numerous bacterial species has enabled the application of very large data sets to phylogeny reconstruction. In this section, we consider the bacterial phylogenies obtained through the analysis of extensive protein sequence alignments. Two main approaches have been applied to unify the information from different molecules into a single tree: (i) concatenation of sequences for the different molecules prior to phylogenetic analysis, or (ii) separate phylogenetic analysis of each molecule and *a posteriori* coding of the information from individual phylogenies into a single “supertree”.

Concatenation-based analyses. In the concatenation approach, each of the molecules in the analysis must be present in all species considered. Therefore, attempts to reconstruct the phylogeny of bacteria through sequence concatenations have employed different subsets of widely distributed proteins. Brown et al. (2001) utilized proteins conserved across species from all domains (Archaea, Bacteria and Eukarya), from which they removed those proteins producing

trees that intermixed species from different domains (which was interpreted as evidence of horizontal transfer). The remaining subset of universally distributed proteins contained 14 information-processing molecules, mostly involved in translation. Other studies have deliberately focused on extensive concatenations of translational proteins (Brochier et al. 2002), and more specifically, of ribosomal proteins (Hansmann and Martin 2000, Wolf et al. 2001, Francino, unpublished). Proteins of the translational apparatus, particularly ribosomal proteins, have long been thought to be less prone to horizontal transfers due to their high level of inter-molecular interactions (*viz.*, the “complexity” hypothesis; Jain et al. 1999). Although cases of horizontal transfer of translational molecules (Doolittle and Handy 1998, Zhaxybayeva and Gogarten 2002), including ribosomal proteins (Brochier et al. 2000, Makarova et al. 2001, Zhaxybayeva and Gogarten 2002) and rRNA (Yap et al. 1999), have been reported, most of these proteins seem to produce phylogenies that are congruent among themselves and with the rRNA tree, suggesting that they belong to a core set of rarely transferred genes (Brochier et al. 2002).

The phylogenies obtained in most concatenation-based analyses retrieve the monophyly of all established bacterial phyla. [This is not the case for Hansmann and Martin (2000), whose analysis contains a likely long-branch attraction artifact between the epsilon-proteobacterium *Helicobacter* and *Mycoplasma*.] In the cases where Chlorobi and Bacteroidetes are represented, they are grouped together with very high bootstrap support (Brown et al. 2001, Brochier et al. 2002). Remarkably, these phylogenies, although not identical, recover several of the same high-level clades. For example, Chlamydiae and Spirochaetes group together in most concatenation-based analyses (Hansmann and Martin 2000, Brown et al. 2001, Wolf et al. 2001, Brochier et al. 2002), a relationship not detected in single-gene phylogenies. Similarly, Aquificae and Thermotogae are often grouped together, and occur at the base in rooted versions

of the bacterial tree (Hansmann and Martin 2000, Wolf et al. 2001, Brochier et al. 2002, Francino, unpublished), although not in the analysis by Brown et al. (2001) where they remain in their usual placement as basal but separate lineages.

Most interesting are groupings that involve deeper relationships between several phyla. A cluster retrieved in several studies is that of Actinobacteria, *Deinococcus*-*Thermus* and Cyanobacteria (Wolf et al. 2001, Brochier et al. 2002, Francino, unpublished), although the sister relationships within the group differ among analyses. Actinobacteria and Cyanobacteria cluster together in the work of Wolf et al. (2001), whereas *Deinococcus* clusters with Actinobacteria in the work of Brochier et al. (2002) and with Cyanobacteria in a more recent analysis with a much wider representation of cyanobacterial species (Francino, unpublished). In the work of Brown et al. (2001), *Deinococcus* also groups with Cyanobacteria, but the relative placement of this cluster is not resolved. In addition, in all three analyses (Wolf et al. 2001, Brochier et al. 2002, Francino, unpublished) the Actinobacteria, *Deinococcus*-*Thermus* and Cyanobacteria cluster at a deeper level with the Firmicutes. Finally, another well-supported deep branch separating two large assemblages of phyla is recovered in the analyses of Brochier et al. (2002) and Francino (unpublished). This branch separates Proteobacteria, Chlamydiae, Spirochaetes, Bacteroidetes and Chlorobi (not present in Francino's analysis) from all other phyla. As mentioned, the HSP60 tree of Viale et al. (1994) also separates Proteobacteria, Chlamydiae, Spirochaetes and Bacteroidetes from the remaining represented phyla (Actinobacteria, Cyanobacteria and Firmicutes), and no large-scale, alignment-based phylogeny has a high level of support for any branch that would contradict this grouping.

Supertree construction. In an alternative approach, Daubin et al. (2001, 2002) propose the independent analysis of individual protein phylogenies, followed by coding of the information obtained from each phylogeny into a matrix that will be used to generate a “supertree”. This approach is less restrictive than sequence concatenation because it can incorporate phylogenies based on different sub-samples of species, and therefore can exploit the information contained in proteins that have limited distribution. In addition, the separate phylogenetic analysis of each protein alignment enables the use of specifically optimized sequence evolution models. Daubin et al. (2002) generated trees for 310 proteins present in at least 10 bacterial species with sequenced genomes, from which they retained a subset of 120 that supported similar phylogenies and were therefore judged to represent a common history, mostly devoid of horizontal transfer. For each protein tree, supported clades were encoded in a binary matrix, with weighting proportional to the bootstrap value, and the individual matrices were concatenated into a single matrix which was used for maximum parsimony tree reconstruction (Baum 1992, Ragan 1992).

The supertree obtained recovers the monophyly of all established bacterial phyla as well as some of the high-level clades supported by other analyses. As with translational proteins (Brochier et al. 2002), *Deinococcus* and Actinobacteria cluster together in the supertree. However, unlike the majority of concatenation-based studies, the supertree does not significantly support the grouping of Cyanobacteria with *Deinococcus* and Actinobacteria, or with any other phylum. The position of these phyla in the bacterial tree is therefore still controversial. Nevertheless, several arguments favor a grouping of *Deinococcus*-*Thermus* with Cyanobacteria. This cluster is recovered in the large-scale studies of Brown et al. (2001) and Francino (unpublished), and in the analysis of a concatenation of SSU and LSU rRNAs by Brochier et al. (2002). Moreover, this relationship is also supported by chaperonin HSP70 trees and by patterns

of indels in several proteins (see below) (Gupta et al. 1997, Gupta 1998, Gupta and Johari 1998). The clustering of *Deinococcus* with Actinobacteria in the supertree and in the analysis of translational proteins by Brochier et al. (2002) may be an artifact due to shared GC-rich base composition. The initial placement of the Deinococcus-Thermus phylum after hyperthermophiles near the base of the bacterial tree is not supported by any large-scale alignment-based studies, nor by later versions of the SSU rRNA tree (Hugengoltz et al. 1998a,b).

The supertree strongly groups the phyla Aquificae and Thermotogae, but there is no support for them as the basal lineage of Bacteria. This result mirrors that of Brochier and Philippe (2002) based on slowly evolving rRNA positions and a well sampled representation of bacterial species. The grouping of the hyperthermophiles, although observed in many recent analyses (Hansmann and Martin 2000, Daubin et al. 2001, 2002, Wolf et al. 2001, Brochier and Philippe 2002, Brochier et al. 2002, Francino, unpublished), may result from biases in nucleotide and amino acid composition due to adaptation to high temperatures. Nevertheless, the recurrence of this grouping in large-scale analyses and the relatively frequent displacement of these phyla from the base of the bacterial tree (Gruber and Bryant 1997, Klenk et al. 1999, Bocchetta et al. 2000, Schutz et al. 2000, Daubin et al. 2001, 2002, Brochier and Philippe 2002) seriously challenge the long held view that thermophily is the ancestral state. Instead of the hyperthermophiles, the supertree favors the Spirochaetes and Chlamydiae as the most basal bacterial lineages. Unlike large-scale concatenation-based analyses (Hansmann and Martin 2000, Brown et al. 2001, Wolf et al. 2001, Brochier et al. 2002), the supertree does not group these two phyla together, but rather places them as separate lineages at the base of the bacterial tree. Other phyla have also been reported to occupy the most basal position in the bacterial tree, most notably the Planctomycetes, in the analysis of slowly evolving rRNA sites by Brochier and Philippe (2002).

Hence, the resolution of the rooting of the bacterial tree still requires a much broader sample of genomic sequences for the different candidate phyla.

Indel signature sequences. In addition to traditional approaches based on the analysis of nucleotide or amino acid substitutions, large sets of protein alignments have also been inspected for insertions and deletions (indels) to be employed as markers for phylogenetic reconstruction. Certain protein sequences are polymorphic for indels, and the presence or absence of particular indels can be used to assort bacteria into groups. If identical insertion or deletion events do not arise independently in different lineages, these characters could also be used to infer the branching order and relationships among organisms. Gupta (2000, 2001) has shown that the occurrence of specific indels defines most bacterial phyla as cohesive groups, and by tracing the distribution of numerous indels in different proteins, he has hypothesized an order of appearance for these phyla. Although the assignment of an ancestral state is not explicit in such reconstructions, the indel “signatures” place the Gram-positive bacteria as closely aligned with the Archaea – satisfying because both have a single-membrane cell wall producing a positive Gram stain– and invoke a single origin for all two-membrane bacteria.

In conclusion, large-scale alignment-based approaches are providing new insights into deep relationships between the bacterial phyla, but different analyses are generating conflicting results, most notably regarding the placement of the root of the bacterial tree (fig.1). Although bacterial genome sequences are now appearing at a rapid rate, species sampling is still highly biased, mostly towards pathogenic and cultivable members of a restricted number of phyla. A further increase in resolution of bacterial relationships awaits a more widespread sampling of lineages. In addition, enhanced resolution will require the development of new methodologies

that enable the optimal integration of information from large numbers of protein alignments. At present, the increased phylogenetic signal detected in large-scale analyses, and the detection of a core of genes that have undergone similar histories (Brochier et al. 2002, Daubin et al. 2002), suggest that there is a phylogenetic framework that reflects the evolution of substantial fractions of bacterial genomes.

LOOKING BEYOND SEQUENCE EVOLUTION

In addition to comparisons of the primary sequences of DNA or proteins, several other molecular, genetic and genomic features have been used to examine the evolutionary relationships among Bacteria (Eisen 2000, Wolf et al. 2002). The goal of such studies is to infer a whole-organism (as opposed to a single gene) phylogeny by developing statistics that denote the overall similarity between entire genomes. To their advantage, these methods do not rely upon the alignments of individual genes or proteins for phylogenetic reconstruction and largely avoid inconsistencies that result from paralogy and lateral transfer unless occurring on a very broad scale. Unfortunately, the rates and patterns of evolution of many of these alternate characters are unclear such that the underlying basis of variation is not known, not specified, or not consistent among taxa. As a result, the reconstructed phylogenies are usually compared with that of SSU rRNA, and any incongruities help model the evolution of a particular trait. For example, if organisms with identical GC contents are unrelated according to their SSU rRNA sequences, the phylogenetic signal of such a character is presumed to be weak and one can conclude that the processes that modulate base composition act somewhat independently of ancestry. The problems (and arguments) arise when discordant phylogenies are inferred from characters that are each assumed to have a common basis in all organisms, to be transmitted

strictly vertically, and to be resistant to convergence (such that similar traits could not have arisen independently in different taxa).

Many whole-genome characteristics can provide measures of similarity among organisms that can be used to infer phylogenetic relationships. Early genetic maps showed that both gene contents and gene order are conserved among closely related organisms (*e.g.*, *Escherichia coli* and *Salmonella typhimurium*) but that the conservation is disrupted with evolutionary distance (*e.g.*, *Escherichia coli* vs. *Bacillus subtilis*). Huynen and Bork (1998) compared the rates of change in gene contents and gene order with that of protein sequence evolution in nine fully sequenced genomes (8 Bacteria and 1 Archaea). Gene order decayed most rapidly, followed by gene repertoire, and finally protein sequences remained most conserved (implying that these possess the strongest phylogenetic signal for inferring relationships among very divergent organisms). Bacterial genomes have been characterized based on numerous other features, and although no explicit phylogenies are produced, these descriptors have been used to infer the intergenomic differences among organisms (Karlin et al. 2002).

Genome content. In a prevalent whole-genome approach, the degree of similarity between organisms is based on the fraction of genes shared. Whereas the mechanisms by which bacterial genomes can gain or lose genes are well understood, the effects of these processes and the rates at which they occur are highly variable across organisms. Because closely related organisms tend to have many uniquely shared characteristics, one anticipates that their gene contents should be similar; however, the estimated fraction of genes common to two organisms can be greatly affected by differences in genome sizes. In the worst case scenario, if an hypothetical organism contained only the minimal set of universally distributed genes, then, based on its gene contents, this genome would be perceived as being equally similar to all other genomes regardless of its

true ancestry. And in the reciprocal comparisons, the fraction of genes shared with this hypothetical organism would be wholly dependent upon the total gene number of a particular organism.

Despite such caveats, many of the whole genome phylogenies based on gene contents resemble those based on SSU rRNA. Naturally, the perceived similarity between genomes is influenced by how shared genes are identified. There are basically two approaches, designated by House and Fitz-Gibbon (2002) as the “homolog” method, which examines the taxonomic distribution of gene families (Fitz-Gibbon and House 1999) or the representation of a genome in the different Clusters of Orthologous Groups of proteins (COGs; Tatusov et al. 1997, Lin and Gerstein 1999, Tatusov et al. 2001, Wolf et al. 2001), and the “ortholog” method, based on the proportion of shared orthologs, usually regarded as sequences that are reciprocally the best matches in a pair of genomes (Snel et al. 1999, Tekaia et al. 1999, Bansal and Meyer 2002, Clarke et al 2002, Korbel et al. 2002). The relationships between genomes by the homology method would not be affected by gene duplication or the partial loss of a gene family because these processes do not alter the absolute number of gene families; but this number is sensitive to massive genome reduction (as seen in many parasites and symbionts) and to the acquisition or loss of novel classes of genes. In contrast, the orthology method would be affected by any mechanism that removes orthologs (deletions, non-homologous gene displacement, etc.) and by recent gene transfers and duplications.

As expected, large genomes have higher numbers of genes or gene families in common (Snel et al. 1999, Lin and Gerstein 2000, Bansal and Meyer 2002, Wolf et al. 2001); but when the metric is normalized by genome size, or when microbes with reduced genomes are excluded from the analysis, the distance between genomes more-or-less reflects traditional phylogenetic

classifications and subdivisions. The gene-content phylogeny of Snel et al (1999) is reminiscent of the SSU phylogeny but groups Cyanobacteria with Aquificae at the base of the bacterial tree, whereas a revised version (Huynen et al. 1999), which applies the same orthology approach but includes additional taxa, places Aquificae alone at the base and the Thermotogae with the Spirochaetes. In contrast, the genome content tree of Clarke et al. (2002), which is based on the fraction of ORFs reaching a normalized BLASTP threshold, places Thermotogae as the basal lineage. The gene content phylogenies based on the homology approach are greatly affected by the inclusion of organisms with highly reduced genomes (which tend to group together) and more “phylogenetically reasonable clades” (Wolf et al. 2001) are recovered by omitting these taxa. Looking at the shared occurrence of gene families in the genomes of sequenced, free-living organisms, House and Fitz-Gibbon (2002) place the Aquificae at the base of the bacterial tree but the branching order of other bacterial phyla remains obscure.

Ortholog similarity. Although different pairs of orthologs conserved between two genomes can display dissimilar levels of identity as a result of differential functional constraints, the distribution of distances between orthologs should reflect the evolutionary relationship between a given pair of genomes. Several authors (Grishin et al. 2000, Clarke et al. 2002, Wolf et al. 2002) have identified orthologs common to two fully sequenced genomes by searching for the reciprocal best hits and used various measures to convert ortholog similarity scores to a matrix conveying the pairwise distance between genomes. Depending on which taxa are included in the particular study, the phylogenies extracted from these distance matrices differ in placing Thermotogae, Aquificae or both at the base of the bacterial tree.

Gene order. Early comparisons of sequenced genomes detected little long-range conservation in gene order (Mushegan and Koonin 1996); but overlaid on this pattern, there are several gene

clusters and gene pairs (mostly involving ribosomal proteins) maintained in all bacterial genomes (Siefert et al. 1997, Dandekar et al. 1998, Wachterhauser 1998, Wolf et al. 2001). Conservation of gene order seems to reflect the evolutionary distance between organisms (Tamames 2000, Wolf et al. 2001), but it evolves faster than gene contents (Huynen and Bork 1998, Korbel et al. 2002), and its use as a phylogenetic marker should be limited to comparisons of rather closely related genomes (Suyama and Bork 2001). Kunisawa (2001) has developed a method for reconstructing phylogenetic relationships based on the physical arrangement of clustered genes, and when applied to members of the Proteobacteria, this method produces a topology that is consistent with the traditional sequence-based classifications.

Other genome features. Because of the well-known heterogeneity in the sizes and nucleotide compositions of bacteria genomes, there have been attempts to group bacterial taxa according to total G+C contents, codon usage biases, di-, tri- and tetra- nucleotide frequencies (genomic signature abundances), the number and distribution of repetitive sequences, short palindromes and rRNA operons, gene and protein lengths, and amino acid composition (reviewed in Karlin et al. 2002). These approaches have retrieved some intriguing (coincidental?) similarities between genomes that are not typically grouped in sequence-alignment-based phylogenies, but few of these features can be viewed as a consistent and reliable indicator of evolutionary distance because many are susceptible to convergence.

Wolf et al. (1999) and Lin and Gerstein (2000) have taken a unique approach by examining the occurrence of specific protein folds throughout the genome, and the phylogenetic groupings in the resulting “fold” trees are surprisingly similar to those based on SSU rRNA. Several authors have begun to make direct comparisons of the tree topologies of Bacteria recovered from different features of the genome and proteome (Lin and Gerstein 2000, Wolf et al. 2001, 2002).

These studies clearly show that some aspects of the genome contribute only weak phylogenetic signals and imply that the addition of characters to an analysis, even those that are known to be genetically determined, will not always increase the accuracy of the tree.

ENTEROBACTERIACEAE: WHAT'S IN A NAME?

Even for the most extensively studied group of Bacteria, recent molecular phylogenetic analyses have revealed some unrecognized kinships between taxa as well as numerous inconsistencies in the present classification scheme. The Enterobacteriaceae are Gram negative, facultatively anaerobic rods belonging to the gamma class of Proteobacteria. Only few actually reside in an “enteric” environment; however, most fall within the group of related species that colonize the mammalian intestine. There are currently 41 described genera in *Bergey's Manual of Systematic Bacteriology* 2nd Edition (2001) including species found in terrestrial and aquatic environments, and associated with plants or animal hosts. Members of this family are usually motile, catalase positive, oxidase negative, and use the Embden-Meyerhof pathway to catabolize sugars and to produce acids from the fermentation of glucose. They can be distinguished from other Gram-negative, rod-shaped bacteria (such as those in the families Vibrionaceae and Pasteurellaceae) by cell geometry, flagellar arrangement, oxidase production, sodium requirements and the presence of a common antigen (Brenner 1981).

Despite these and other shared metabolic properties, assignment of microorganisms to the Enterobacteriaceae no longer requires detailed information about the morphological and biochemical characteristics of an organism. Molecular phylogenetic analyses have now defined the extent of sequence variation within this taxonomic group, and any strain whose sequence falls within this delineated range can be tentatively classified to the Enterobacteriaceae, even

without knowledge of its morphology, metabolic capabilities or growth requirements. Of course, this assumes that the phylogenetic trees and evolutionary relationships based on genes represent those of the organisms themselves; and at least for many of the species constituting the Enterobacteriaceae, the correspondence between the molecular and the original phenotypic classification schemes is rather good. That said, we should note that molecular analyses have revealed that several enteric genera are polyphyletic, such that their constituent species are not derived from a unique common ancestor. For example, some species of *Klebsiella* and *Enterobacter* are more closely related to one another than to other species typed to the same genus (Harada and Ishikawa 1997, Mollet et al. 1997, Hauben et al. 1998, Dauga 2002). Such findings have resulted in the recent reclassification of certain taxa and changes in nomenclature. At present, the determination of a comprehensive phylogeny encompassing all genera within a family can not yet rely on whole-genome sequencing, and relationships at this and lower taxonomic levels are still determined through single-gene phylogenies, mostly SSU rRNA.

Phylogenetic analyses of Enterobacteriaceae based on SSU rRNA. The phylogenetic tree presented in figure 2 displays the evolutionary relationships of many well-recognized genera of Enterobacteriaceae and, for reference, *Haemophilus influenzae* and *Vibrio cholerae*, which are non-enteric members of the gamma-Proteobacteria. This phylogeny is derived from SSU rRNA sequences; and, in many cases, relationships among the better-studied organisms match those resolved by biotyping. It is evident that certain genera are cohesive and monophyletic, that is, contained within a group that contains a single common ancestor and all of its descendent organisms. As mentioned above and as previously reported (Hauben et al. 1998, Sproer et al. 1999, Dauga 2002), *Klebsiella*, *Enterobacter*, and *Serratia* are polyphyletic, consisting of species descended from two or more ancestors.

To construct the phylogeny in figure 2, SSU rRNA sequences for representatives of all recognized enteric genera (except *Moellerella*, *Saccharobacter*, *Trabulsiella* and *Yokenella* for which cognate sequences are not available) were extracted from GenBank. For several genera, numerous sequences are available (either from multiple species within the genus or from multiple rRNA operons within a genome) yielding a set of 193 enteric SSU rRNA sequences, which were aligned with ClustalX using the default settings (Thompson et al. 1997). Following initial alignment, sequences were removed if they were not full length or if the genera and/or species were not well represented in the literature. Next, we removed all but one (or two) sequences from the set for each genus if the multiple representatives for a given genus were found to be monophyletic. The resulting data set contained 27 taxa, including 22 enteric genera (Table 1). The NEXUS file containing these sequences is available at TreeBase (<http://www.treebase.org>) under Study Accession S801 and Matrix Accession M1267.

To refine the sequence data for phylogenetic analysis, we excluded two short regions (positions 70-103 and 461-489) due to high levels of sequence and length variation, which made alignments ambiguous. To determine the best-fit model of DNA evolution, the resulting alignments were subjected to hierarchical likelihood ratio tests in Modeltest v3.06 (Posada and Crandall 1998). For the phylogenetic analyses, a neighbor-joining (NJ) tree was constructed under the general time reversible (GTR) model of evolution (log likelihood = -6952.9580; estimated by Modeltest v3.06) using PAUP* v4.0b10 (Swofford 2000). The maximum parsimony (MP) analysis was conducted using PAUP* by optimizing characters with accelerated transformation, 10 repetitions of random sequence additions, starting trees obtained by stepwise addition and branches swapped by tree bisection-reconstruction. In addition, a Bayesian inference of phylogeny was conducted using MrBayes v2.01 (Huelsenbeck and Ronquist 2001).

The GTR model, with some sites assumed to be invariant and with variable sites assumed to follow a discrete gamma distribution, was used for all analyses. Bayesian analyses were initiated with random starting trees and run for 1.5 million generations, with a 100 generations sampling interval. Replicates were performed to determine that analyses were not trapped at local optima (Huelsenbeck and Bollback 2001). After discarding the initial burn-in samples (the first 5,000 trees), the remaining 10,000 trees were used to generate a 50% majority rule consensus tree using PAUP*, with the percentage of tree containing a node representing that node's posterior probability (Huelsenbeck and Ronquist 2001).

Figure 2 represents the most reliable composite estimation of the phylogenetic relationships among enteric bacteria. A time scale for the diversification of bacterial species has been proposed (Ochman and Wilson 1987, Doolittle et al. 1996) based on the estimated divergence time of 100 million years ago for the split between *Escherichia coli* and *Salmonella enterica*. If SSU rRNA evolves at approximately the same rate in all enteric taxa (Ochman et al. 1999), then it is possible to estimate the divergence time between any pair of taxa by extrapolation.

Structural RNAs evolve very slowly and are not useful for resolving the relationships among closely related organisms, such as strains within a species or, sometimes, even species within a genus. In addition to SSU rRNA, several other molecules and typing methods have been used to examine relationships among enterics, and below, we briefly discuss the phylogenetic status of species within representative enteric genera:

Escherichia/Shigella. The genus *Escherichia* contains five species: *Escherichia blattae*, *E. coli*, *E. fergusonii*, *E. hermannii*, and *E. vulneris* whose relationships to one another, and to other enteric species, have been resolved by the phylogenetic analysis of *gapA* and *ompA* sequences

(Lawrence et al. 1991). *E. coli* and *E. fergusonii* form a monophyletic clade, with each being more closely related to one another than either is to *Salmonella*; however the other three species classified to this genus represent a phylogenetically heterogeneous group of organisms. *E. hermanii* is basal to *E. coli*, *Salmonella enterica* and *Citrobacter freundii*; *E. blattae* branches between *Serratia* and *Klebsiella*, and *E. vulneris* is polyphyletic, with one strain grouping with *K. pneumoniae*.

The relationships of thousands of strains of *E. coli* have been determined by multilocus enzyme electrophoresis, and several other typing schemes have examined the genetic variation in clinical, laboratory, and natural isolates (Whittam 1997). It is probably safe to say that more is known about the degree, appointment and maintenance of genetic variation in *E. coli* than in any other bacterial species. The genera *Escherichia* and *Shigella* were grouped together by DNA-DNA hybridization studies, which demonstrated that they belonged to the same genetic species (Brenner et al. 1972, 1973). Subsequent analyses have shown that the variation within the four recognized species of *Shigella* (*S. dysenteriae*, *S. flexneri*, *S. boydii* and *S. sonnei*) is contained within that observed in *E. coli*, i.e., strains of *Shigella* have arisen through multiple independent origins from within *E. coli* (Ochman et al. 1983, Pupo et al. 1997, 2000).

Salmonella. The genus *Salmonella* is monophyletic and comprises two species, *Salmonella bongori* and *S. enterica*. The separation of *S. bongori* from *S. enterica* has been confirmed by DNA-DNA hybridization (Le Minor et al. 1982) and phylogenetic analyses of housekeeping, and rRNA genes (Selander et al. 1996, Christensen et al. 1998). *S. bongori* is considered to be a homogeneous group (Hoszowski and Wasyl 2000) whereas *S. enterica* is subdivided into seven subspecies, enumerated as I, II, IIIa, IIIb, IV, VI and VII, which encompass more than 2,000 serovars. Most human pathogens (e.g., Typhi) belong to subspecies I. The relationships within

Salmonella have been derived from multilocus enzyme electrophoresis and by the analysis of several genes, which have been sequenced in each of the subspecies type strains constituting the SARC collection (Boyd et al. 1996, Brown et al. 2002)

Erwinia/Pectobacterium/Brenneria/Pantoea. Originally, the genus *Erwinia* was divided into two taxonomic groups, *amylovora* and *carotovora*, based on phytopathogenicity properties (Brown et al. 2000). The *amylovora* group is considered less metabolically active and requires organic nitrogen for growth, whereas the *carotovora* group has higher metabolic activity, reduces nitrates to nitrites, and causes soft-rot diseases (Brenner 1981). Subsequently, phylogenetic analyses of SSU rRNA demonstrated that the genus *Erwinia* was polyphyletic, which led to the partitioning of *Erwinia* into three genera: *Erwinia*, *Pectobacterium* and *Brenneria* (Hauben et al. 1998). The genera *Erwinia* and *Brenneria* contain the necrogenic phytopathogens (formerly the *amylovora* group) and *Pectobacterium* encompasses the soft-rotting phytopathogens (formerly the *carotovora*). An additional group of phytopathogens closely related to *Erwinia* has been assigned to the genus *Pantoea*.

Serratia. Members of the genus *Serratia* can be found in environmental samples as well as in association with human disease. There are nine described species within the genus: *Serratia odorifera*, *S. marcescens*, *S. rubidaea*, *S. entomophila*, *S. ficaria*, *S. fonticola*, *S. plymuthica*, *S. proteamaculans* and *S. grimesii* (Sproer et al. 1999). Phylogenetic analyses of SSU rRNA sequences suggest that *Serratia* is polyphyletic, with one cluster more closely related to *Hafnia alvei* and *Obesumbacterium proteus*, than to the other *Serratia* cluster (Sproer et al. 1999). In contrast, phylogenetic analyses of the sequences encoding the ATPase domain of DNA gyrase (*gyrB*) showed *Serratia* to contain two clusters forming a monophyletic group (Dauga 2002).

Klebsiella. Klebsiellae are detected in environmental samples as well as on the mucosal surfaces of mammals. Certain species are opportunistic pathogens and are associated with respiratory diseases and urinary tract infections in humans (Podschun and Ullman 1998). The five most common species are *Klebsiella aerogenes*, *K. edwardsii*, *K. pneumoniae*, *K. ozaenae* and *K. rhinoscleromatis*, and in some classifications, *K. edwardsii* and *K. pneumoniae* are subdivided into additional subspecies (Podschun and Ullman 1998). Phylogenetic analyses of SSU rRNA (Hauben et al. 1998), translation initiation factor 2 (*infB*; Hedegaard et al. 1999), RNA polymerase β -subunit (*rpoB*; Mollet et al. 1997), *groE* (Harada and Ishikawa 1997) and *gyrB* (Dauga 2002) all suggest that the genus is polyphyletic. In these studies, some species of *Klebsiella* were more closely related to members of the genus *Enterobacter* than to other *Klebsiella* species (Hauben et al. 1998, Dauga 2002). In light of this evidence, there is a need to reclassify the members of this genus in a way that reflects their molecular relationships.

Enterobacter. The *Enterobacter* occupy similar environmental and ecological niches as *Klebsiella*. There are 16 described species within the genus, and as with *Klebsiella*, the genus *Enterobacter* is polyphyletic. Based on the phylogenetic analyses of nucleic acid and protein-coding sequences, certain *Enterobacter* species should be grouped with *Klebsiella*. For example, *E. aerogenes* falls in a clade with *Klebsiella pneumoniae* rather than with *E. cloacae* in phylogenies based on SSU rRNA and *gyrB* sequences (Dauga 2002).

Other enteric genera. Some of the other members of the *Enterobacteriaceae* exhibit a wide-range of associations with eukaryotic hosts, ranging from opportunistic pathogens to mutualistic symbionts. Some examples of pathogens include *Edwardsiella*, which are pathogenic to fish, certain *Yersinia* and *Citrobacter*, which cause diarrhea and disease in humans, and *Proteus*, which induce otitis in dogs. In contrast, *Photorhabdus* and *Xenorhabdus* consist almost entirely

of mutualistic symbionts of the nematode genera *Steinernema* and *Heterorhabditis* (Boemare et al. 1993). These symbionts serve to maintain conditions conducive for nematode reproduction (Poinar and Thomas 1966, 1967) and produce antimicrobial compounds that inhibit the growth of other microorganisms (Akhurst 1983, Boemare et al. 1993).

The phylogenetic relationships within several of these enteric genera have been analyzed using SSU rRNA sequences [*e.g.*, *Buttiauxella* (Sproer et al. 1999), *Photorhabdus* and *Xenorhabdus* (Suzuki et al. 1996, Liu et al. 1997), and *Yersinia* (Ibrahim et al 1993)]. And for those enteric genera containing but a single species [*e.g.*, *Calymmatobacterium* (Kharsany et al. 1999), *Candidatus* (Zreik et al. 1998) and *Hafnia* (Hauben et al. 1998, Sproer et al. 1999), their relationships to other members of the Enterobacteriaceae (as well as their identification as unique genera) has typically been established by the analysis of SSU rRNA sequences.

THE EMERGENCE OF BACTERIAL SYMBIONTS

Among the greatest impact of culture-independent methods for the identification and classification of microbes has been the characterization of bacterial symbionts, which are broadly distributed and can not be maintained outside of their animal hosts. Insects that feed exclusively on blood or plant sap contain obligate (“primary”) symbionts that offset their host nutritional deficiencies, reside within specialized host organs and are essential for host survival. In addition to these primary symbionts, many insects contain “secondary” symbionts that confer no obvious benefit to their host, are not essential for either host growth or reproduction, and invade a diverse range of host tissues. Whereas primary symbionts are strictly vertically transmitted within a host lineage, sometimes for hundreds of millions of years (as evident from the concordance between symbiont and host phylogenies), secondary symbionts can be horizontally transmitted among insect host species (Moran and Baumann 2000).

Molecular phylogenetic analyses have provided evidence that the majority of primary and secondary symbionts of insects have evolved from within the Proteobacteria, and in particular the gamma-Proteobacteria (fig. 3). To show the relationships of symbiotic lineages to the Enterobacteriaceae, we included several enteric species in this phylogeny and rooted this tree with *Vibrio cholerae*, which also serves as an outgroup to the enteric bacteria (fig. 2).

The evolution of symbionts from free-living proteobacterial ancestors often involves a vast reduction in genome size; in fact, at 440 kb, the bacterial symbiont of aphids has the smallest known genome for a cellular life form (Gil et al. 2002). As is apparent from this phylogeny, the origin of new symbiotic species can occur by two routes: (i) the independent derivation of symbionts from free-living ancestors, and (ii) the diversification from within a clade that is already symbiotic (usually associated with the infection of a new host species). The molecular genetic information for most of the symbionts included in fig. 3 is presently limited to SSU rRNA sequences; however, their small genome sizes and limited gene repertoires make many of these bacteria ideal subjects for genome projects.

ACKNOWLEDGEMENTS

The authors acknowledge the generous support of the Department of Energy project number DEF0301ER63147. We are also grateful to Jake Russell and Nancy Moran for their input.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

LITERATURE CITED

- Akhurst RJ (1983) Taxonomic study of *Xenorhabdus*, a genus of bacteria symbiotically associated with insect pathogenic nematodes. *Int J Syst Bacteriol* 33: 38-45
- Baldauf SL, Palmer JD, Doolittle WF (1996) The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc Natl Acad Sci USA* 93: 7749-7754
- Bansal AK, Meyer TE (2002) Evolutionary analysis by whole-genome comparisons. *J Bacteriol* 184: 2260-2272
- Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3-10
- Bocchetta M, Gribaldo S, Sanangelantoni A, Cammarano P (2000) Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. *J Molec Evol* 50:366-380
- Boemare NE, Akhurst RJ, Mourant RG (1993) DNA relatedness between *Xenorhabdus* spp (*Enterobacteriaceae*), symbiotic bacteria of entomopathogenic nematodes, and a proposal to transfer *Xenorhabdus luminescens* to a new genus, *Photorhabdus* gen nov. *Int J Syst Bacteriol* 43: 249-255
- Boyd EF, Wang F-S, Whittam TS, Selander RK (1996) Molecular relationships of the salmonellae. *Appl Env Microbiol* 62: 804-808
- Brenner DJ (1981) Introduction to the family *Enterobacteriaceae*. In Starr MP, Stolp H, Trüper HG, Balows A, Schegel H (eds) *The Prokaryotes: a Handbook on Habitats, Isolation and Identification of Bacteria*. Springer-Verlag, Berlin, pp 1105-1127
- Brenner DJ, Fanning GR, Skerman FJ, Falkow S (1972) Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *J Bacteriol* 109: 9593-9605

- Brenner DJ, Fanning GR, Miklos GV, Steigerwalt, AG (1973) Polynucleotide sequence relatedness among *Shigella* species. *Int J Syst Microbiol* 23: 1-7
- Brenner DJ, Steigerwalt AG, Gail Wathen H, Gross RJ, Rowe B (1982) Confirmation of aerogenic strains of *Shigella boydii* 13 and further study of *Shigella* serotypes by DNA relatedness. *J Clin Microbiol* 16: 432-436
- Brochier C, Philippe H (2002) A non-hyperthermophilic ancestor for Bacteria. *Nature* 417: 244
- Brochier C, Philippe H, Moreira D (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet* 16: 529-533
- Brochier C, Baptiste E, Moreira D, Philippe H (2002) Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* 18: 1-5
- Brown EW, Davis, RM, Gouk, C, van der Zwet T (2000) Phylogenetic relationships of necrogenic *Erwinia* and *Brenneria* species as revealed by glyceraldehyde-3-phosphate dehydrogenase gene sequences. *Int J Syst Evol Microbiol* 50: 2057-2068
- Brown EW, Kotewicz ML, Cebula T (2002) Detection of recombination among *Salmonella enterica* strains using the incongruence length difference test. *Molec Phylogenet Evol* 24: 102-120
- Brown JR, Masuchi Y, Robb FT, Doolittle WF (1994) Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. *J Molec Evol* 38:566-576
- Brown JR, Doolittle WF (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci USA* 92: 2441-2445
- Brown JR, Doolittle WF (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol Molec Biol Rev* 61: 456-489

- Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. *Nat Genet* 28:281-285
- Cavalier-Smith T (1992) Origins of secondary metabolism. *Ciba Found Symp* 171: 64-80
- Cavalier-Smith T (2002) The neomuran origin of archaeobacteria, the negibacterial root of the universal tree and bacterial megaclassification. *Int J Syst Evol Microbiol* 52: 7-76
- Chang HR, Loo LH, Jeyaseelan K, Earnest L, Stackebrandt E (1997) Phylogenetic relationships of *Salmonella typhi* and *Salmonella taphimurium* based on 16S rDNA sequence analysis. *Int J Syst Bacteriol* 47: 1253-1254
- Christensen H, Olsen JE (1998) Phylogenetic relationships of *Salmonella* based on DNA sequence comparison of *atpD* encoding the β subunit of ATP synthase. *FEMS Microbiol Letters* 161: 89-96
- Christensen H, Nordentoft S, Olsen JE (1998) Phylogenetic relationship of *Salmonella* based on ribosomal RNA sequences. *Int J Syst Bacteriol* 48: 605-610
- Clarke GDP, Beiko RG, Ragan MA, Charlebois RL (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol* 184: 2072-2080
- Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* 23: 324-328
- Daubin V, Gouy M, Perriere G (2001) Bacterial molecular phylogeny using supertree approach. *Genome Inform Ser Workshop Genome Inform* 12: 155-164
- Daubin V, Gouy M, Perriere G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* 12: 1080-1090

- Dauga, C (2002) Evolution of the *gyrB* gene and the molecular phylogeny of *Enterobacteriaceae*: a model molecule for molecular systematic studies. *Int J Syst Evol Microbiol* 52: 531-547
- Desjardins P, Picard B, Kaltenbock B, Elion J, Denamur E (1995) Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. *J Molec Evol* 40: 440-448
- De Rijk P, Van de Peer Y, Van den Broeck I, De Wachter R (1995) Evolution according to large ribosomal subunit RNA. *J Molec Evol* 41:366-375
- Delwiche CF, Kuhsel M, Palmer JD (1995) Phylogenetic analysis of *tufA* sequences indicates a cyanobacterial origin of all plastids. *Molec Phylogenet Evol* 4:110-128
- Dodd CER, Jones D (1982) A numerical taxonomic study of the genus *Shigella*. *J Gener Microbiol* 128: 1933-1957
- Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science* 284: 2124-2129
- Doolittle WF, Handy J (1998) Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr Opin Genet Dev* 8:630-636
- Doolittle RF, Feng DF, Tsang S, Cho G, Little E (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271: 470-477
- Eisen JA (1995) The RecA protein as a model for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J Molec Evol* 41: 1105-1123
- Eisen JA (1998) A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res* 26: 4291-4300

- Eisen JA (2000) Assessing evolutionary relationships among microbes from whole-genome analysis. *Curr Opin Microbiol* 3: 475-480
- Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res* 27: 4218-4222
- Forterre P, Phillipe H (1999) Where is the root of the universal tree of life? *Bioessays* 21: 871-879
- Fox GE, Magnum LJ, Balch WE, Wolfe RS, Woese CR (1977) Classification of methanogenic bacteria by 16S ribosomal RNA characterization. *Proc Natl Acad Sci USA* 74: 4537-4541
- Fox GE and 18 co-authors (1980) The phylogeny of prokaryotes. *Science* 209: 457-463
- Gil R, Sabater-Muñoz B, Latorre A, Silva FJ, Moya A (2002) Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. *Proc Natl Acad Sci USA* 99: 4454-4458
- Gogarten JP (1995) The early evolution of cellular life. *Trends Ecol Evol* 10:147-151
- Gogarten JP and 12 co-authors (1989) Evolution of the vacuolar H⁺-ATPase: implications for the origins of eukaryotes. *Proc Natl Acad Sci USA* 86: 6661-6665
- Golding GB, Gupta RS (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Molec Biol Evol* 12: 1-6
- Gruber TM, Bryant DA (1997) Molecular systematic studies of eubacteria, using sigma 70-type sigma factors of group 1 and group 2. *J Bacteriol* 179: 1734-1747
- Grishin NV, Wolf YI, Koonin EV (2000) From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res* 10: 991-1000

- Gruber TM, Eisen JA, Gish K, Bryant DA (1998) The phylogenetic relationships of *Chlorobium tepidum* and *Chloroflexus aurantiacus* based upon their RecA sequences. *FEMS Microbiol Lett* 162:53-60
- Gupta RS (1997) Protein phylogenies and signature sequences: evolutionary relationships within prokaryotes and between prokaryotes and eukaryotes. *Antonie Van Leeuwenhoek* 72: 49-61
- Gupta RS (1998) Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes. *Microbiol Molec Biol Rev* 62: 1435-1491
- Gupta RS (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiol Lett* 24: 367-402
- Gupta RS (2001) The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins. *Int Microbiol* 4: 187-202
- Gupta RS, Johari V (1998) Signature sequences in diverse proteins provide evidence of a close evolutionary relationship between the *Deinococcus-Thermus* group and cyanobacteria. *J Molec Evol* 46:716-720
- Gupta RS, Bustard K, Falah M, Singh D (1997) Sequencing of heat shock protein 70 (DnaK) homologs from *Deinococcus proteolyticus* and *Thermomicrobium roseum* and their integration in a protein-based phylogeny of prokaryotes. *J Bacteriol* 179:354-357
- Gupta RS, Mukhtar T, Singh B (1999) Evolutionary relationships among photosynthetic prokaryotes (*Heliobacterium chlorum*, *Chloroflexus aurantiacus*, cyanobacteria, *Chlorobium tepidum* and proteobacteria): implications regarding the origin of photosynthesis. *Molec Microbiol* 32:893-906

- Hansmann S, Martin (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient cluster that is conserved across prokaryotic genomes: Influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol* 50: 1655-1663
- Harada H, Ishikawa H (1997) Phylogenetical relationship based on *groE* genes among phenotypically related *Enterobacter*, *Pantoea*, *Klebsiella*, *Serratia* and *Erwinia* species. *J Gen Appl Microbiol* 43: 355-361
- Hauben L, Moore ERB, Vauterin L, Steenackers M, Mergaert J, Verdonck L, Swings J (1998) Phylogenetic position of phytopathogens within the *Enterobacteriaceae*. *System Appl Microbiol* 21: 384-397
- Hedegaard J, Steffensen SA, Norskov-Lauritsen N, Mortensen KK, Sperling-Petersen HU (1999) Identification of *Enterobacteriaceae* by partial sequencing of the gene encoding translation initiation factor 2. *Int J Syst Bacteriol* 49: 1531-1538
- Hilario E, Gogarten JP (1993) Horizontal transfer of ATPase genes – the tree of life becomes a net. *Biosystems* 31: 111-119
- Hillis DM, Dixon MT (1991) Ribosomal RNA: molecular evolution and phylogenetic inference. *Quart Rev Biol* 66: 411-453
- Hoszowski A, Wasyl D (2000) Taxonomy and nomenclature of genus *Salmonella*. *Medycyna Weterynaryjna* 56: 75-78
- House CH, Fitz-Gibbon ST (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J Molec Evol* 54: 539-547
- Huelsenbeck JP, Bollback JP (2001) Empirical and hierarchical Bayesian estimation of ancestral states. *Syst Biol* 50: 351-366

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogeny.

Bioinformatics 17: 754-755

Hugenholtz P, Pitulle C, Hershberger K, Pace NR (1998a) Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol* 180: 366-376

Hugenholtz P, Goebel BM, Pace NR (1998b) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* 180: 4765-4474

Huynen M, Gutell R, Konings D (1997) Assessing the reliability of RNA folding using statistical mechanics. *J Molec Biol* 267: 1104-1112

Huynen MA, Bork P (1998) Measuring genome evolution. *Proc Natl Acad Sci USA* 95: 5849-5856

Huynen M, Snel B, Bork P (1999) Lateral gene transfer, genome surveys, and the phylogeny of prokaryotes. *Science* 286: 1443a

Ibrahim A, Goebel BM, Liesack, W, Griffiths M, Stackebrandt, E (1993) The phylogeny of the genus *Yersinia* based on 16S rDNA sequences. *FEMS Microbiol Letters* 114: 173-178

Iwabe N, Kuma K, Hasegawa H, Osawa S, Miyata AT (1989) Evolutionary relationship of Archaea, Bacteria, eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA* 86: 9355-9359

Jain R, Rivera MC, Lake JA (1999) Horizontal transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci USA* 96: 3801-3806

Jenkins C, Fuerst JA (2001) Phylogenetic analysis of evolutionary relationships of the Planctomycete division of the domain Bacteria based on amino acid sequences of elongation factor Tu. *J Molec Evol* 52:405-418

- Karlin S, Weinstock GM, Brendel V (1995) Bacterial classifications derived from RecA protein sequence comparisons. *J Bacteriol* 177: 6881-6893
- Karlin S, Brocchieri L, Trent J, Blaisdell B, Mrazek J (2002) Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. *Theor Pop Biol* 61: 367-390
- Kharsany ABM, Hoosen AA, Kiepiela P, Kirby R, Sturm AW (1999) Phylogenetic analysis of *Calymmatobacterium granulomatis* based on 16S rDNA gene sequences. *J Med Microbiol* 48: 841-847
- Kjer KM (1995) Use of rRNA secondary structure in phylogentic studies to identify homologous positions: an example of alignment and data presentation from the frogs. *Molec Phylogenet Evol* 4: 314-330
- Klenk H-P, Meier T-D, Durovic P, Schwass V, Lottspeich F, Dennis PP, Zillig W (1999) RNA polymerase of *Aquifex pyrophilus*: implications for the evolution of the bacterial rpoBC operon and extremely thermophilic bacteria. *J Molec Evol* 48:528-541
- Korbel JO, Snel B, Huynen MA, Bork P (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet* 18: 158-162
- Kunisawa T (2001) Gene arrangements and phylogeny in the class Proteobacteria. *J Theor Biol* 213: 9-19
- Lawrence JG, Ochman H., Hartl DL (1991) Molecular and evolutionary relationships among enteric bacteria. *J Gener Microbiol* 137: 1911-1921
- LeMinor L, Veron M, Popoff M (1982) Taxonomie des *Salmonella*. *Ann Microbiol Inst Pasteur* 133B: 223-243
- Lin J, Gerstein M (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res* 10: 808-818

- Liu J, Berry R, Poinar G, Moldenke A (1997) Phylogeny of *Photorhabdus* and *Xenorhabdus* species and strains as determined by comparison of partial 16S rDNA gene sequences. *Int J Syst Bacteriol* 47: 948-951
- Ludwig W, Neumaier J, Klugbauer N, Brockmann E, Roller C, Jilg S, Reetz K, Schachtner I, Ludvigsen A, Bachleitner M, Fischer U, Schleifer KH (1993) Phylogenetic relationships of Bacteria based on comparative sequence analysis of elongation factor Tu and ATP-synthase beta-subunit genes. *Antonie Van Leeuwenhoek* 64:285-305
- Ludwig W, Strunk O, Klugbauer S, Klugbauer N, Weizenegger M, Neumaier J, Bachleitner M, Schleifer KH (1998) Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* 19: 554-568
- Ludwig W, Klenk H-P (2001) Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In Garrity GM (ed), *Bergey's manual of systematic bacteriology 2nd edition, vol. 1 The Archaea and the Deeply Branching and Phototrophic Bacteria*. Springer-Verlag, NY
- Maidak BL and 12 co-authors (2000) The RDP (Ribosomal Database Project) continues. *Nucleic Acids Res* 28: 173-174
- Makarova KS, Ponomarev VA, Koonin EV (2001) Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol* 2: RESEARCH 0033.1-0033.14
- Mollet C, Drancourt M, Raoult D (1997) *rpoB* sequence analysis as a novel basis for bacterial identification. *Molec Microbiol* 26: 1005-1011
- Moran NA, Baumann P (2000) Bacterial endosymbionts in animals. *Curr Opin Microbiol* 3: 270-275

- Mushegian AR, Koonin EV (1996) Gene order is not conserved in bacterial evolution. *Trends Genet* 12: 289-290
- Ochman, H, Whittam TS, Caugant DA, Selander RK (1983) Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. *J Gen Microbiol* 129: 2715-2729
- Ochman H, Wilson AC (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J Molec Evol* 26: 74-86
- Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. *Proc Natl Acad Sci USA* 96: 12638-12643
- Pace NR (1991) Origin of life – facing up to the physical setting. *Cell* 65:531-533
- Philippe H, Laurent J (1998) How good are deep phylogenetic trees? *Curr Opin Genet Dev* 8:616-623
- Podschun R, Ullman U (1998) *Klebsiella* spp as nosocomial pathogens: Epidemiology, taxonomy, typing methods and pathogenicity factors. *Clin Microbiol Rev* 11: 589-603
- Poinar GO, Thomas GM (1966) Significance of *Achromobacter nematophilus* Poinar and Thomas (Archromobacteraceae: Eubacteriales) in the development of the nematode DD-136 (*Neoaplectana* sp : Steinernematidae). *Parasitology* 56: 385-390
- Poinar GO, Thomas GM (1967) The nature of *Achromobacter nematophilus* as an insect pathogen. *J Invert Pathol* 9: 510-514
- Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* 14: 817-818

- Pupo GM, Karaolis DKR, Lan R, Reeves PR (1997) Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. *Infect Immun* 65: 2685-2692
- Pupo GM, Lan R, Reeves PR (2002) Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci USA* 97: 10567-10572
- Ragan MA (1992) Phylogenetic inference based on matrix representation of trees. *Molec Phylogenet Evol* 1:53-58
- Schutz M and 11 others (2000) Early evolution of cytochrome bc complexes. *J Molec Biol* 300: 663-675
- Selander RK, Li J, Nelson K (1996) Evolutionary genetics of *Salmonella enterica*. In Neidhardt FC (ed) *E coli and Salmonella: Cellular and Molecular Biology* Vol 2, ASM Press, Washington, D C, pp 2691-2702
- Siefert JL, Martin KA, Abdi F, Widger WR, Fox GE (1997) Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J Molec Evol* 45: 467-472
- Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. *Nat Genet* 21: 108-110
- Sproer C, Mendrock U, Swiderski J, Lang E, Stackebrandt E (1999) The phylogenetic position of *Serratia*, *Buttiauxella* and some other genera of the family *Enterobacteriaceae*. *Int J Syst Bacteriol* 49: 1433-1438
- Suyama M, Bork P (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet* 17: 10-13

- Suzuki T, Yabusaki H, Nishimura Y (1996) Phylogenetic relationships of entomopathogenic nematophilic bacteria: *Xenorhabdus* spp. and *Photorhabdus* spp. *J Basic Microbiol* 5: 351-354
- Swofford, DL (2000) PAUP* *Phylogenetic Analysis Using Parsimony* v. 4. Sinauer, Sunderland, MA
- Tamames J (2001) Evolution of gene order conservation in prokaryotes. *Genome Biol* 2: RESEARCH0020
- Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28: 33-36
- Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278: 631-637
- Teichmann SA, Mitchison G (1999) Is there a phylogenetic signal in prokaryote proteins? *J Molec Evol* 49: 98-107
- Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res* 9: 550-557
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignments aided by quality analysis tools. *Nucleic Acids Res* 24: 4876-4882
- Tiboni O, Cammarano P, Sanangelantoni AM (1993) Cloning and sequencing of the gene encoding glutamine synthetase I from the archaeum *Pyrococcus woesei*: anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthetase I sequences. *J Bacteriol* 175:2961-2969

- Viale AM, Arakaki AK, Soncini FC, Ferreyra RG (1994) Evolutionary relationships among eubacterial groups as inferred from GroEL (Chaperonin) sequence comparisons. *Int J Syst Bacteriol* 44:527-533
- Wachtershauser G (1998) Towards a reconstruction of ancestral genomes by gene cluster alignment. *Syst Appl Microbiol* 21: 473-477
- Whittam TS (1996) Genetic variation and evolutionary processes in natural populations of *Escherichia coli*. In Neidhardt FC (ed) *E. coli and Salmonella: Cellular and Molecular Biology* Vol 2, ASM Press, Washington, DC, pp 2708-2720
- Winker S, Woese CR (1991) A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics. *Syst Appl Microbiol* 14: 305-310
- Woese CR (1987) Bacterial evolution. *Microbiol Rev* 51: 221-271
- Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87:4576-4579.
- Wolf YI, Brenner SE, Bash PA, Koonin EV (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Res* 9: 17-26
- Wolf YI, Aravind L, Grishin NV, Koonin EV (1999) Evolution of aminoacyl-tRNA synthetases – analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res* 9: 689-710
- Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* 1: 8
- Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. *Trends Genet* 18: 472-479

- Yap WH, Zhang Z, Wang Y (1999) Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol* 181:5201-5209
- Zhaxybayeva O, Gogarten JP (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses. *BMC Genomics* 3: 4
- Zreik L, Bove, JM, Garnier M (1998) Phylogenetic characterization of the bacterium-like organism associated with marginal chlorosis of strawberry and proposition of a *Candidatus* taxon for the organism, '*Candidatus* Phlomobacter fragariae'. *Int J Syst Bacteriol* 48: 257-261

Table 1: Sample information for *Enterobacteriaceae* sequences used in the phylogenetic analyses.

Bacterial Species	Stain	Accession no.
<i>Brenneria (Edwinia) salicis</i>	DSM 30166	AJ233419
<i>Buchnera aphidicola</i>	strn. Sg (<i>Schizaphis graminum</i>)	NC_004061 ^a
<i>Citrobacter freundii</i>	DSM 30039	AJ233408
<i>Edwardsiella tarda</i>	NB8031	AB050832
<i>Enterobacter cloacae</i>	CCCO10	AF511434
<i>Enterobacter intermedius</i>	ATCC 33110	AF310217
<i>Erwinia amylovora</i>	BC205	AF141896
<i>Escherichia coli</i>	O157:H7	NC_002695 ^a
<i>Haemophilus influenzae</i>	KW20	NC_000907 ^a
<i>Hafnia alvei</i>	ATCC 13337	M59155
<i>Klebsiella planticola</i>	ATCC 33531T	AF129443
<i>Klebsiella pneumoniae</i>	AU1496	AY043391
<i>Morganella morganii</i>	M567	AF461011
<i>Pantoea stewartii</i>	DC283	AJ311838
<i>Pectobacterium (Erwinia) chrysanthemi</i>	582	AF373175

<i>Photorhabdus asymbiotica</i>	ATCC 43950	Z76755
<i>Plesiomonas shigelloides</i>	ATCC 7966T	X60418
<i>Proteus vulgaris</i>	ATCC 29906T	AJ301683
<i>Providencia stuartii</i>	ATCC 29914T	AF008581
<i>Rahnella aquatilis</i>	CDC 1327-79	U90757
<i>Salmonella bongori</i>	BR1859	AF029227
<i>Salmonella enterica</i> sv. <i>Typhi</i>	CT18	NC_003198 ^a
<i>Serratia grimesii</i>	LMG 7883	AF286868
<i>Serratia marcescens</i>	AU1388	AY043389
<i>Shigella dysenteriae</i>	ATCC 13313	X96966
<i>Vibrio cholera</i>	ATCC 14033	X74694
<i>Xenorhabdus nematophila</i>	CB6	AF522294
<i>Yersinia enterocolitica</i>	ATCC 9610	AF366378
<i>Yersinia pestis</i>	CO92	NC_003143 ^a

^a Genbank accession number of completely sequenced genome

FIGURE LEGENDS

Figure 1. Relationships among bacterial phyla most often represented in large-scale phylogenetic analyses. The cladogram depicts the basic topology of the SSU rRNA tree: Aquificae and Thermotogae are placed at the base, and all other phyla appear in a “bush”-like zone, where only a clade uniting Bacteroidetes and Chlorobi is resolved. Colored boxes indicate relationships strongly supported by other types of phylogenetic analyses, mostly based on protein sequence concatenations or whole-genome characters. The orange arrow indicates the proposed association between Aquificae and Proteobacteria. Purple arrows indicate alternative candidate phyla for the most basal position in the tree. The phylum Planctomycetes has no representatives for which a complete genome is available and has not been included in analyses based on extensive data sets. See text for references to the studies supporting each of the depicted relationships.

Figure 2. Phylogeny of Enterobacteriaceae based on SSU rRNA. See text for criteria used to select species and for methods of phylogenetic reconstruction.

Figure 3. SSU rRNA phylogeny of gamma-proteobacterial symbionts and related free-living species. PE, primary endosymbiont; SS, secondary symbiont. Many symbionts have no official nomenclature and are listed according to their designations in the literature or in sequence databases. Common names and latin binomes of hosts are given.

?

Planctomycetes