Overview: A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics

Wolfgang Ludwig and Hans-Peter Klenk

Introduction

Despite its relatively short history, microbial systematics has never been static but rather constantly subject to change. The evidence of this change is provided by many reclassifications in which bacterial taxa have been created, emended, or dissected, and organisms renamed or transferred. The development of a procaryotic systematics that reflects the natural relationships between microorganisms has always been a fundamental goal of taxonomists. However, the task of elucidating these relationships could not be addressed until the development of molecular methods (the analysis of macromolecules) that could be applied to bacterial identification and classification. Determination of genomic DNA G + C content, and chemotaxonomic methods such as analysis of cell wall and lipid composition, in many cases proved superior to classical methods based upon morphological and physiological traits. These tools provide information that can be used to differentiate taxa, but do not allow a comprehensive insight into the genetic and phylogenetic relationships of the organisms. DNA–DNA reassociation techniques provide data on genomic similarity and hence indirect phylogenetic information, but the resolution of this approach is limited to closely related strains. DNA–DNA hybridization is the method of choice for delimiting procaryotic species and estimating phylogeny at and below the species level. The current species concept is based on two organisms sharing a DNA–DNA hybridization value of greater than 70% (Wayne et al., 1987).

With improvement in molecular sequencing techniques, the idea of Zuckerkandl and Pauling (1965) to deduce the phylogenetic history of organisms by comparing the primary structures of macromolecules became applicable. The first molecules to be analyzed for this purpose were cytochromes and ferredoxins (Fitch and Marguliash, 1967). Subsequently, Carl Woese and coworkers demonstrated the usefulness of small subunit (SSU) rRNA as a universal phylogenetic marker (Fox et al., 1977). These studies suggested natural relationships between microorganisms on which a new procaryotic systematics could be based. The aims of this chapter are to provide a brief description of the methods used to reconstruct these phylogenetic relationships, to explore the phylogenetic relationships suggested by 16S rRNA and alternative molecular chronometers, and to present a justification for the use of the current 16S rRNA-based procaryotic systematics as a backbone for the structuring of the second edition of *Bergey's Manual of Systematic Bacteriology*.

Reconstruction and Interpretation of Phylogenetic Trees

Sequence alignment The critical initial step of sequence-based phylogenetic analyses is undoubtedly the alignment of primary structures. Alignment is necessary because only changes at positions with a common ancestry can be used to infer phylogenetic conclusions. These homologous positions have to be recognized and arranged in common columns to create an alignment, which then provides the basis for subsequent calculations and conclusions. Sequences such as SSU rRNA that contain a number of conserved sequence positions and stretches can be aligned using multiple sequence alignment software such as CLUSTAL W (Swofford et al., 1996). Furthermore, these conserved islands can be used a guide for arranging the intervening variable regions. The alignment of variable regions may remain difficult if deletions or insertions have occurred during the course of evolution. In addition, the homologous character of positions in variable regions is not necessarily indicated by sequence identity or similarity and hence can often not be reliably recognized. However, functional homology, if detectable or predictable, can be used to improve the alignment. In the case of rRNAs, functional pressure apparently dictates the evolutionary preservation of a common core of secondary or higher order structure which is manifested by the potential participation of 67% of the residues in helix formation by intramolecular base pairing. The majority of these structural elements are identical or similar with respect to their position within the molecule as well as number and position of paired bases, or internal and terminal loops. The primary structure sequence alignment can be evaluated and improved by checking for potential higher structure formation (Ludwig and Schleifer, 1994). Furthermore, the character of the base pairing, G–C versus non-G–C, Watson–Crick versus non-Watson–Crick, may be used to refine an alignment. The pairing is a byproduct of thermodynamic stability and consequently has an impact on function. Therefore, adjustments to the alignment appear rational from an evolutionary point of view. However, the recognition of homologous positions in regions which are highly variable with respect to primary as well as higher order structure may still be difficult or even impossible.

The principal problems of aligning rRNA sequences can be avoided by the routine user, if they take advantage of comprehensive databases of aligned sequences (including higher order structure information) that can be obtained from the Ribosomal Database Project (Maidak et al., 1999), the compilations of small

TABLE 1. Transformation of measured distances (lower triangle) into phylogenetic distances (upper triangle): applying the Jukes Cantor (Jukes and Cantor, 1969) transformation[a]

	Escherichia coli	*Klebsiella pneumoniae*	*Proteus vulgaris*	*Pseudomonas aeruginosa*	*Bacillus subtilis*	*Thermus thermophilus*	*Geotoga subterranea*
Escherichia coli		3.2	7	15.6	26	28.5	35.8
Klebsiella pneumoniae	3.1		7	15.1	25.8	28.2	36.4
Proteus vulgaris	6.7	6.7		17.6	26.6	29.9	37.8
Pseudomonas aeruginosa	14.1	13.7	15.7		23.5	29.2	34.3
Bacillus subtilis	22	21.8	22.4	20.2		27	30.4
Thermus thermophilus	23.7	23.5	24.7	24.2	22.6		32.4
Geotoga subterranea	28.5	28.8	29.7	27.6	25	26.3	

[a]The uncorrected distances were used for the reconstruction of the tree in Fig. 1. Given that the data are not ultrametric (see Swofford et al, 1996), they do not directly correlate with the branch lengths in the tree.

and large subunit rRNAs at the University of Antwerp (De Rijk et al., 1999; Van de Peer et al., 1999), or the ARB project as a guide to inserting new sequence data. The RDP offers alignment of submitted sequences as a service while the ARB program package contains tools for automated alignment, secondary structure check, and confidence test.*

Treeing methods The number and character of positional differences between aligned sequences are the basis for the inference of relationships. These primary data are then processed using treeing algorithms based on models of evolution. Usually, the phylogenetic analysis is refined by positional selection or weighting according to criteria such as variability or likelihood. The results of these analyses are usually visualized as additive trees. Terminal (the "organisms") and internal (the common "ancestors") nodes are connected by branches. The branching pattern indicates the path of evolution and the (additive) lengths of peripheral and internal branches connecting two terminal nodes indicate the phylogenetic distances between the respective organisms. There are two principal versions of presentation: radial trees or dendrograms (Fig. 1). The advantage of radial tree presentation is that phylogenetic relationships, especially of only moderately related groups, can usually be shown more clearly, and that all of the information is condensed into an area which can be inspected "at a glance". However, the number of terminal nodes (sequences, organisms, taxa) for which the relationships can be demonstrated is limited. This number is not limited in dendrograms.

A number of different sequence databased treeing methods or algorithms have been developed. Most of them are based on models of evolution. These models describe assumed rules of the evolutionary process concerning parameters such as (overall) base frequencies or (number and weighting of) substitution types. A comprehensive review on methods for phylogenetic analyses, models of evolution, and the mathematical background is given by Swofford et al. (1996). The three most commonly used treeing methods, distance matrix, maximum parsimony, and maximum likelihood, operate by selecting trees which maximize the congruency of topology and branch lengths with the measured data under the criteria of a given model of evolution.

Distance treeing methods such as Neighbor Joining (Saitou and Nei, 1987) or the method of Fitch and Margoliash (Fitch and Margoliash, 1967) rely on matrices of distance values obtained by binary comparison of aligned sequences and calculation of the fraction of base differences. These treeing programs mostly perform modified cluster analyses by defining pairs and, subsequently, groups of sequences sharing the lowest distance values and connecting them into the framework of a growing tree. The tree topology is optimized by maximizing the congruence between the branch lengths in the tree and the corresponding inferred distances of the underlying matrix.

Before treeing, the measured differences are usually transformed into evolutionary distance values according to models of evolution. The underlying assumption is that the real number of evolutionary changes is underestimated by counting the detectable differences in present day sequences. For example, the Jukes Cantor transformation (Jukes and Cantor, 1969) accounts for this underestimation by superelevation of the measured distances (Table 1). Although the theoretical assumptions that provide the basis for transforming the measured distance values into phylogenetic distances are convincing with respect to overall branch lengths, there is a certain risk of misinterpretation or overestimation of local tree topologies. An intrinsic disadvantage of distance treeing methods is that only part of the phylogenetic information, the distances, is used, while the character of change is not taken into account. However, there are methods available to perform more sophisticated distance calculations than simply counting the differences (Felsenstein, 1982).

In contrast to distance methods, maximum parsimony-based treeing approaches use the original sequence data as input. According to maximum parsimony criteria, tree reconstruction and optimization is based on a model of evolution that assumes preservation to be more likely than change. Parsimony methods search for tree topologies that minimize the total tree length. That means the most parsimonious (Edgell et al., 1996) tree topology (topologies) require(s) the assumption of a minimum number of base changes to correlate the tree topology and the original sequence data. In principle, the problem of plesiomorphies (see below) can be handled more appropriately with parsimony than with distance methods, given that the most probable ancestor character state is estimated at any internal node of the tree. Long branch attraction is a disadvantage of the maximum parsimony approach. The parsimony approach does infer branching patterns but does not calculate branch lengths *per se.* To superimpose branch lengths on the most parsimonious tree topologies additional methods and criteria have to be applied. Both PAUP* and ARB parsimony tools are able to combine the

**Editorial Note:* Software available from O. Strunk and W. Ludwig, Department of Microbiology, Technische Universität, München, Munich, Germany. ARB is a software environment for sequence data located at: www.mikro.biologie.tu-muenchen.de/pub/ARB.

**Editorial Note:* Software available from David Swofford at the Laboratory of Molecular Systematics, National Museum of Natural History, Smithonian Institution, Washington, D.C. Available from Sinauer Associates of Sunderland, MS at www.sinauer.com/formpurch.htm.

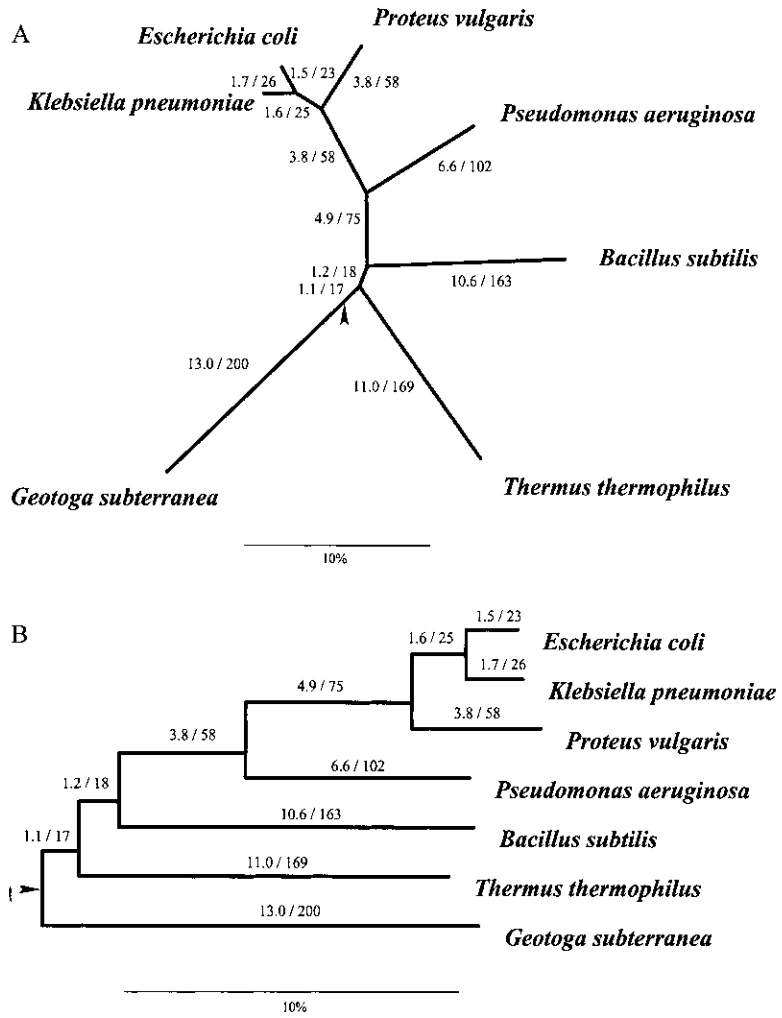

FIGURE 1. Additive trees. The same tree is shown as a radial tree (*A*) and a dendrogram (*B*). The tree was reconstructed by applying the neighbor joining method (Saitou and Nei, 1987) to a matrix of uncorrected binary 16S rRNA sequence differences for the organisms shown in the tree and a selection of archaeal sequences as outgroup references. *Arrowheads* indicate the branching of the archaeal reference sequences and the root of the trees. Bar = 10% sequence difference. The distance between two sequences (organisms) is the sum of all branch lengths directly connecting the respective terminal nodes or the sum of the corresponding horizontal branch lengths in the radial tree or the dendrogram, respectively. The numbers at the individual branches indicate overall percentage sequence divergence, followed by the number of different sequence positions (the length of the *E. coli* 16S rRNA sequence [1542 nucleotides] was used as reference in all calculations). Note: the tree topology was not evaluated by applying different methods and parameters.

reconstruction of topologies and the estimation of branch lengths.

The most sophisticated of the three independent phylogenetic treeing methods is maximum likelihood, where a tree topology is regarded as optimal if it reflects a path of evolution that, according to the criteria of given models of evolution, most likely resulted in the sequences of the contemporary organisms. The corresponding evolutionary models may include parameters such as transition/transversion ratio, positional variability, character state probability per position and many others. Given that the maximum likelihood approach utilizes more of the information content of the underlying sequences, it is considered to be superior to the other two treeing methods. An accompanying disadvantage is the need for expensive computing time and performance. Even if powerful computing facilities are accessible only a limited number of sequences can be handled within a reasonable time. Rapid development in the field of computing hardware suggests that this powerful method may become applicable for larger data sets in the near future.

The use of filters Most commonly used programs for phylogenetic treeing are capable of including filters or weighting masks that remove or weight down individual alignment columns while treeing, thus reducing the influence of highly variable positions. Conservation profiles can be calculated by simply determining the fraction of the most frequent character. More sophisticated approaches define positional variability, the rate of change, or the likelihood of a given character state, with respect to an underlying tree topology according to parsimony criteria, or by using a maximum likelihood approach. The choice of phylogenetic entities for which filters or masks should be generated depends on the group of organisms or the phylogenetic level

TABLE 2. Phylogenetic information content of procaryotic small subunit rRNA[a]

	Bacteria				*Archaea*			
Intra-domain similarity	&mt;67%				&mt;67%			
	Conserved		Variable		Conserved		Variable	
	Pos.	%	Pos.	%	Pos.	%	Pos.	%
Sequence conservation	568	36.8	974	63.2	571	37	971	63
Potential information (bits)	1948				1942			
Number of characters	1	2	3	4	1	2	3	4
Positional variability, %	36.8	23.2	13.5	26.5	37	28.3	15.2	19.5
Corrected information (bits)	1506				1385			

[a]The calculations were performed using the 16S rRNA sequence of *E. coli* (1542 nucleotides) as a reference. To avoid influences of sequencing, database, and alignment errors a 98% similarity criterion was applied to define "invariant" positions. Therefore, the term 'conserved' was used instead "invariant". Bits (of information) were calculated by multiplying the logarithm to the base two of the permissive character states (positional variability: different nucleotides per position) times the number of informative (variable) sites. Potential information was calculated as the maximum information content assuming positional variability of four. These values were corrected according measured positional variability.

(the corresponding area in a tree) of interest. Tools for the generation of profiles, masks, or filters are implemented in the ARB software package or available from other authors (Swofford et al., 1996; Maidak et al., 1999). The removal of positions also means loss of information; therefore it is recommended to perform treeing analyses of a given data set several times applying different filters. This helps to visualize the robustness or weakness of a specific tree topology and to estimate whether or not variable positions have had a substantial influence. Filters or masks should only be calculated using comprehensive data sets of full sequences; then these filters can also be applied to the analysis of partial sequences. The results of many years of tree reconstruction have shown that positions should only be removed up to 60% positional conservation, to avoid the loss of too much information. In most cases use of a 50% conservation filter is appropriate.

Confidence tests Different treeing methods handle data according to particular assumptions and consequently may yield different results. The many inconsistencies of real sequence data also prevent easy and reliable phylogenetic inference; therefore the careful evaluation of tree topologies is to be recommended. Besides the application of filters and weighting masks and the use of different treeing approaches, resampling techniques can be used to evaluate the statistical significance of branching order. Bootstrapping or jackknifing (Swofford et al., 1996) are procedures that randomly sample or delete columns in sequence data (alignments) or distance values (distance matrix). Usually 100–1000 different artificial data sets are generated as inputs for treeing operations by these methods. For each data set the optimum tree topologies are inferred by the particular treeing method and, finally, a consensus tree topology is generated. In this consensus tree, bootstrap or jackknife values are assigned to the individual branches. These values indicate the number of treeing runs in which the subtree defined by the respective branch appeared as monophyletic with respect to all other groups. An example of a bootstrapped tree is shown in Fig. 2. Besides the bootstrap value, an area of low significance is indicated by circles centered on the individual (internal) nodes. These areas were estimated from the sampling values in relation to the corresponding (internal) branch lengths using the ARB software tools. No convincing significance can be expected if only a few residues provide information supporting the separation of branches or subtrees. Given that in most cases branch lengths indicate the degree of estimated sequence divergence, a subtree separated from the remainder of a phylogenetic tree by a short internal branch is highly unlikely to be assigned a high resampling value.

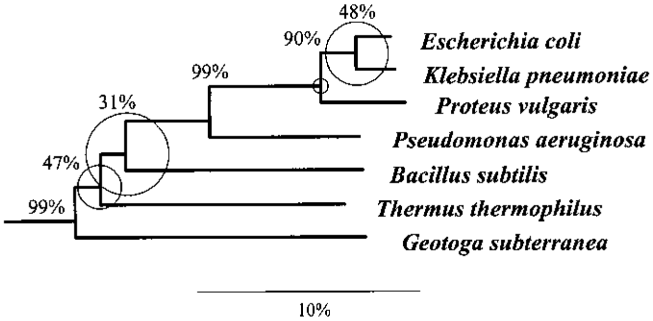

FIGURE 2. Confidence tests on tree topology. 1000 bootstrap operations were performed for evaluation of the tree in Figure 1B. The numbers at the furcations indicate the fraction of (1000 bootstrapped) trees which support the separation of the respective subtree (branches to the right of the particular furcation) from all other branches or groups in the tree. Circles indicating an area of "unsharpness" were calculated as a function of bootstrap values and branch lengths using ARB.

The resampling techniques can only be used to estimate the robustness of a tree reconstructed by applying a single treeing method and parameter set. Thus for reliable phylogenetic conclusions it is necessary to combine different treeing methods, as well as filters and weighting masks and resampling techniques. Even if appropriate software and powerful hardware are accessible, high quality tree evaluations may get rather expensive in working and computing time. An approach for estimating "upper bootstrap" limits without the need of expensive multiple treeings was developed and implemented in the ARB package. In the absence of resampling values, a critical "reading" of trees allows a rough estimate of the confidence of relative branching orders at a glance, assuming that a short branch length in most cases also indicates low significance of separation.

Why do trees differ? Tree reconstruction can often be a frustrating experience, especially for researchers not familiar with the theoretical principles of phylogenetic treeing, when the application of different treeing methods or parameters to a single data set results in different tree topologies. This is not surprising since different treeing methods are based on different models of evolution, and therefore the data are processed in different ways. Consequently, a perfect match of tree topologies cannot necessarily be expected even if identical data sets are analyzed using identical parameters. None of the models reflect perfectly the reality of the evolutionary process. The assumption of independent evolution of different sequence positions, for example, does not hold true for the many functionally correlated residues such as base paired nucleotides in rRNAs. In addition, none of the treeing methods and software programs can really exhaustively test and optimize all possible tree topologies. For example, with only 20 sequences there would be 10^{20} possible tree topologies to be examined. Other factors, such as data selection (the organisms and sequence positions included in calculations), the order of data addition to the tree, and the presence of positions that have changed at a higher rate than the remainder of the data set, also influence tree topology. These instabilities do not usually concern the global tree topology but rather local branching patterns.

Limitations of Tree Reconstruction

Information content of molecular chronometers The reconstruction of gene or organismal history, based upon the degree of divergence of present day sequences, relies on the number and character of detectable sequence changes that have accumulated during the course of evolution. Thus the maximum information content of molecules is defined by the number of characters (monomers), and the number of potential character states (different residues), per site. With real data, only a fraction of the sites are informative, as a reasonable degree of sequence conservation is needed to demonstrate the homologous character of molecules or genes and to recognize a phylogenetic marker as such. For example, there are 974 (63.2%) variable and hence informative positions in the 16S rRNA genes of members of the *Bacteria*, and 971 (63%) such positions in the *Archaea*. Given that the maximum information content per position is defined by the number of possible character states i.e., the four nucleotides (the potential fifth character state, deletion or insertion, is not considered here), there could be 1948 (*Bacteria*) or 1942 (*Archaea*) bits of information (logarithm to base 2 of the number of possible character states times the number of informative positions) in the SSU rRNA. However, due to functional constraints and evolutionary selective pressure, the number of allowed character states varies from position to position. As shown in Table 2, there are only 407 (26.4%; *Bacteria*) or 301 (19.5%; *Archaea*) positions in the investigated data set at which all four nucleotides are found, whereas only three different residues apparently are tolerated at 209 (13.6%; *Bacteria*) or 233 (15.2%; *Archaea*) positions, and only two character states are realized at 358 (23.2%; *Bacteria*) or 437 (28.3%; *Archaea*) positions. Thus the theoretical information content of 1984 (*Bacteria*) or 1938 (*Archaea*) bits in reality is reduced to 1506 (*Bacteria*) or 1385 (*Archaea*). The reduced information content draws attention to the need for careful sequence alignment and analysis.

The problem of plesiomorphy Any homologous residue in present day sequences can only report one evolutionary event. The higher the number of permitted characters at a particular position, the higher the probability that such an evolutionary event is directly detectable (by a difference). The majority of these events remain obscure since, especially at variable positions, identical residues are probably the result of multiple changes during the course of evolution, simulating an unchanged position (plesiomorphy). The effect of plesiomorphy on the topology of the resulting trees depends on the number of plesiomorphies supporting branch attraction and also on the treeing method used. Such plesiomorphic sites may cause misleading branch attraction, as shown in Fig. 3, where a short stretch of aligned real 16S rRNA sequences is used to visualize branch attraction. Plesiomorphies may also be responsible for the observation that long "naked" branches represented by only one or a few highly similar sequences often "jump" in phylogenetic trees when the reference data set is changed or expanded. The positioning can usually be stabilized when further representatives of different phylogenetic levels of that branch become available. The rooting of trees may also be influenced by identities at plesiomorphic sites when single sequences are used as outgroup references. The influence of plesiomorphic positions can be reduced by using them at a lower weight for tree reconstruction, but is nevertheless still present.

Partial sequence data There are several convincing arguments for the use of only complete sequence data in the reconstruction of phylogenetic trees. These include the limited information content of the molecule, and the fact that different parts of the primary structure carry information for different phylogenetic levels (Ludwig et al., 1998b). Whenever partial sequences are added to a database of complete primary structures and phylogenetic treeing approaches are applied to the new data set, the new sequences may influence the overall tree topology. The inclusion of partial sequence data may impair phylogenetic trees or influence conclusions previously based on full data. Software which allows the addition of new data to a given data set, and placement of the new sequence according to optimality criteria in a validated tree without changing its topology, is now available. The ARB implementation of this software is capable of removing short partial sequences from a tree prior to the integration of a new highly similar but more complete sequence. Thus the more informative sequence is not "attracted" by a probably misplaced partial sequence. After finding the most similar sequences, the ARB tool compares the number of determined characters, removes the shorter version, and reinserts the data in the order of completeness. There are a number of recent publications presenting comprehensive trees based upon data sets which have been truncated to the regions comprised by included partial sequences. This procedure is not acceptable, given all the limitations of partial sequence data and of the methods of analysis.

Partial sequence data of appropriate regions of the gene may contain sufficient information for the identification of organisms. The determination and comparative analysis of partial sequences may be sufficient to reliably assign an organism to a phylogenetic group if the database contains sequences from closest relatives. A fraction of the 5′-terminal region of the SSU rRNA (*Escherichia coli* pos. 60–110) is one the most informative or discriminating regions for closely related organisms. Hence partial sequence data that include this region can be used to find the closest relative of an organism or to indicate a novel species. Short diagnostic regions (15–20 nucleotides) of partial sequences can also be used as targets for taxon-specific probes or PCR primers that are commonly used for the sensitive detection and identification of microorganisms (Schleifer et al., 1993; Amann et al., 1995; Ludwig et al., 1998a).

Bush-like trees The majority of names and definitions of major phylogenetic groups, such as the phylum *"Proteobacteria"* and the corresponding classes (*"Alphaproteobacteria"*, *"Betaproteobacteria"*, *"Gammaproteobacteria"*, *"Deltaproteobacteria"*, and *"Epsilonproteobacteria"*) originated in the early years of comparative rRNA sequence analysis. At that time phylogenetic clusters could easily be delimited, given that the trees contained many long "naked"

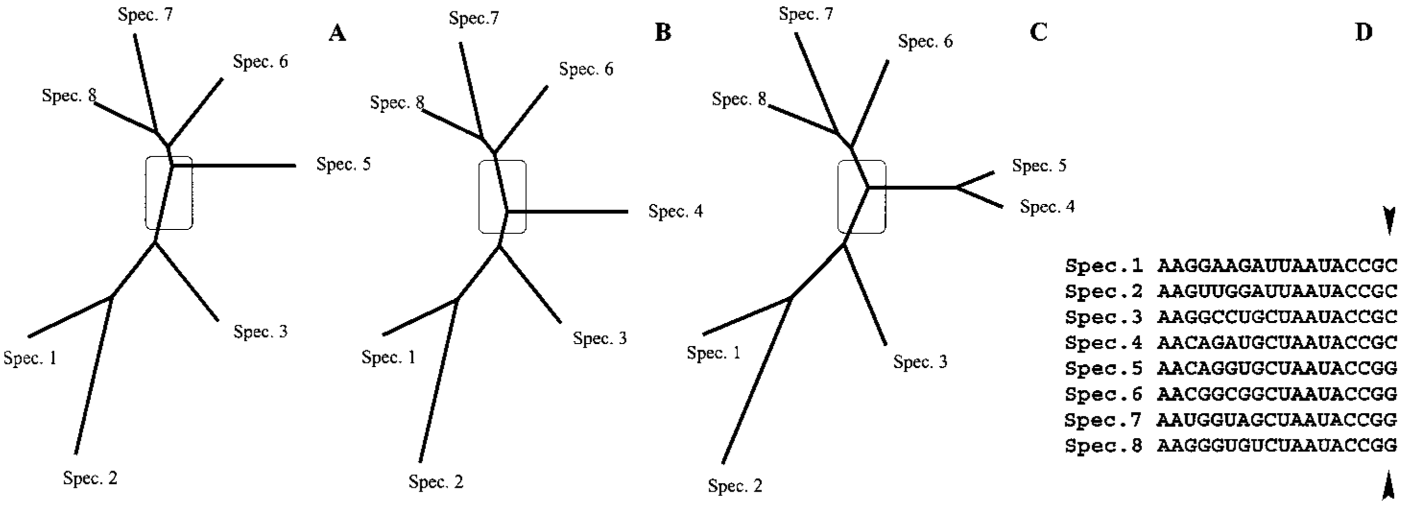
A
B
C
D
Spec. 1
Spec. 2
Spec. 3
Spec. 4
Spec. 5
Spec. 6
Spec.7
Spec. 7
Spec. 8
Spec.1 AAGGAAGAUUAAUACCGC
Spec.2 AAGUUGGAUUAAUACCGC
Spec.3 AAGGCCUGCUAAUACCGC
Spec.4 AACAGAUGCUAAUACCGC
Spec.5 AACAGGUGCUAAUACCGG
Spec.6 AACGGCGGCUAAUACCGG
Spec.7 AAUGGUAGCUAAUACCGG
Spec.8 AAGGGUGUCUAAUACCGG

FIGURE 3. "Branch attraction". The trees show the effects of separate (*A*, *B*) or combined (*C*) inclusion of sequences Spec. 4 and Spec. 5 on the tree topology. The *rectangle* highlights the region of the tree where major changes can be seen. The trees were reconstructed using the neighbor joining method on the aligned 16S rRNA sequence fragments shown (*D*). The column of residues responsible for the attraction of Spec. 5 branch and the Spec. 6–8 subtree in *A* as well as the attraction of Spec. 4 branch and the Spec. 1–3 subtree in *B* is marked by arrows.

branches separating subtrees. This "phylogenetic clarity" was mainly an effect of the limited amount of available sequence data and has often been obscured by the rapid expansion in the number of sequence database entries. Most of the long "naked" branches have expanded and the tree-like topology changed to a bush-like topology. It is probably only a matter of time before the missing links will be found for the remaining "naked" branches such as the *Chlamydiales*, the *"Flexistipes"*, or branches assigned to cloned environmental sequences.

In bush-like areas of a tree, the probability that a given branch will exchange positions with a neighboring branch decreases with the distance between the two branches. This indicates that the relative order of closely neighboring branches cannot be reliably reconstructed, although their separation from more distantly located lineages remains robust. As a consequence, delimitation of taxa often cannot be based on individual local branching order. The use of criteria or additional data for the definition of taxonomic units remains the subjective decision of the taxonomist. In some cases, this leads to the definition of taxa that include paraphyletic groups.

Presentation of Phylogenetic Trees

The main purpose of drawing trees is to visualize the phylogenetic relationships of the organisms or markers, and to allow the reader to recognize these relationships at a glance. It is often difficult to combine an easy-to-grasp presentation of phylogenetic relationships and associated information on the significance of branching patterns. There is no optimum solution to the problem of "correct" presentation of trees; however, ways of addressing this problem can be suggested.

One acceptable procedure would be to present all the trees (which may differ locally) obtained from the same data set by the application of different treeing methods and parameters. However, this may prove more confusing than helpful for readers not experienced in phylogenetic treeing. A more user-friendly solution is to present only one tree topology and to indicate the significance of the individual branches or nodes. However, showing multiple confidence values at individual nodes, or depicting areas of confidence by shading or circles around the nodes, may make the tree unreadable, especially in areas of bush-like topology.

In many cases use of a consensus tree is advantageous. Some programs for consensus tree generation are able to present local topologies as multifurcations at which a relative branching order is not significantly supported by the results of tree evaluations. A fairly acceptable compromise is to use a consensus tree, and to visualize both a detailed branching pattern where stable topologies can be validated, and multifurcations that indicate inconsistencies or uncertainties. Such a multifurcation indicates missing information on that particular era of evolution rather than multiple events resulting in a high diversification within a narrow span of evolutionary time. This type of presentation is certainly more informative for the reader than a choice of various tree topologies, each showing low statistical significance (Ludwig et al., 1998b).

None of the modes of presentation described above can be applied to bush-like topologies and yield meaningful results. Although individual branches are likely to change their positions only locally within a large bush-like area, depending on the methods and parameters applied for treeing, there is no way to split up such an area by several multifurcations. No methods are currently available for the calculation of confidence values for the next, second, third and so on neighboring nodes, and highlighting areas of unsharpness makes the tree difficult to read. A legitimate solution is to base the calculations on the full data set but to hide some of the branches for presentation purposes, and show a tree topology containing a smaller number of significantly separated branches. Thus, while the tree would be based on all available information, only that part of the tree topology of interest for the particular phylogenetic problem is shown clearly laid out (Ludwig et al., 1998b).

16S rRNA: The Benchmark Molecule for Procaryote Systematics

In principle, all the requirements of a phylogenetic marker molecule are fulfilled in SSU rRNAs to a greater extent than in almost all other described phylogenetic markers (Woese, 1987; Olsen and Woese, 1993; Olsen et al., 1994b; Ludwig and Schleifer, 1994; Ludwig et al., 1998b). Besides functional constancy, ubiquitous distribution, and large size (information content), genes coding for SSU rRNA exhibit both evolutionarily conserved regions and highly variable structural elements. The latter characteristic results from different functional selective pressures acting upon the independent structural elements. This varying degree of sequence conservation allows reconstruction of phylogenies for a broad range of relationships from the domain to the species level. A comprehensive SSU rRNA sequence data set (currently more than 16,000 entries) is available in public databases (Ludwig and Strunk, 1995*; Maidak et al., 1999; Van de Peer et al., 1999) in plain or processed (aligned) format, and is rapidly increasing in size. A significant fraction of validly described procaryotic species are represented by 16S rRNA sequences from type strains or closely related strains.

As with any new technique in the field of taxonomy, it took time to establish comparative sequencing of SSU rRNA (genes) as a powerful standard method for the identification of microorganisms and defining or restructuring procaryotic taxa according to their natural relationships. Rapid progress in sequencing and *in vitro* nucleic acid amplification technology led to the replacement of an expensive, sophisticated, and tedious methodology, available only to specialists, by rapid and easy-to-apply routine techniques. As a result, analysis of the genes coding for SSU rRNA is one of the most widely used classification techniques in procaryotic identification and systematics. It is widely accepted that SSU rRNA analysis should be integrated into a polyphasic approach for the new description of bacterial species or higher taxa.

Some Drawbacks of 16S rRNA Gene Sequence Analysis

Functional constraints Depending on functional importance, the individual structural elements of rRNAs cannot be freely changed. It is therefore assumed that sequence change in the rRNAs occurs in jumps rather than as a continuous process. The divergence of present day rRNA sequences may document the succession of common ancestors and their present day descendants, but a direct correlation to a time scale cannot be postulated.

Multiple genes It has been known since the early days of comparative rRNA sequence analysis that the genomes of microorganisms may contain multiple copies of some genes or operons. However, until recently it was commonly assumed that there are no remarkable differences between the rRNA gene sequences of a given organism. A significant degree of sequence divergence among multiple homologous genes within the same organism, such as has been found in *Clostridium paradoxum* (Rainey et al., 1996) and *Paenibacillus polymyxa* (Nübel et al., 1996), would call any sequence-based interorganism relationships into question. The underestimation of this problem may be attributed to the fact that such differences are not easy to recognize using sequencing techniques which depend on purified rRNA or amplified rDNA, and can be mistaken for artifacts. Only frame shifts resulting from inserted or deleted residues can be readily recognized. New techniques, such as denaturing gradient gel electrophoresis (DGGE) (Nübel et al., 1996), allow sequence variation in PCR-amplified rDNA fragments to be detected. The rapidly progressing genome sequencing projects have also provided detailed information on the topic of intraorganism rRNA heterogeneities. Different organisms vary with respect to the presence and degree of intercistron primary structure variation, and most differences concern variable positions and affect base-paired positions (Engel, 1999; Nübel et al., 1996). Although some projects to systematically investigate interoperon differences have been initiated, no comprehensive survey of the spectrum of microbial phyla has been performed. Current and future investigations will show whether regularities or hot spots for interoperon differences can be defined in general or in particular for certain phylogenetic groups. This knowledge can then be used to remove or weight such positions for phylogenetic reconstructions.

Interpretation of high 16S rRNA gene sequence similarity Organisms sharing identical SSU rRNA sequences may be more diverged at the whole genome level than others which contain rRNAs differing at a few variable positions. This has been shown by comparison of 16S rRNA sequence and genomic DNA–DNA hybridization data (Stackebrandt and Goebel, 1994). In the interpretation of phylogenetic trees, it is important to note that branching patterns at the periphery of the tree cannot reliably reflect phylogenetic reality. Given the low phylogenetic resolving power at these levels of close relatedness (above 97% similarity), it is highly recommended to support conclusions based on SSU rRNA sequence data analysis by genomic DNA reassociation studies (Stackebrandt and Goebel, 1994).

Comparative Analyses of Alternative Phylogenetic Markers

Other genes have been investigated as potential alternative phylogenetic markers, to determine whether SSU rRNA-based phylogenetic conclusions can describe the relationships of the organisms, or merely reflect the evolutionary history of the respective genes. For sound testing of phylogenetic conclusions based on SSU rRNA data, the sequences used must originate from adequate phylogenetic markers. The principal requirements for such markers are ubiquitous distribution in the living world combined with functional constancy, sufficient information content, and a sequence database which represents diverse organisms, containing at least members of the major groups (phyla and lower taxa) as defined based upon SSU rRNA.

How many alternative phylogenetic markers are out there? Comparative analysis of the completed genome sequences suggests that there are only a limited number of genes that occur in all genomes and which also share sufficient sequence similarity to be recognized as ortho- or paralogous. Analysis of the first eight completely sequenced genomes (six *Bacteria*, one *Archaea*, and one yeast) showed that only 110 clusters of orthologous groups (COGs) were present in all genomes (Tatusov et al., 1997; Koonin et al., 1998; updated in www.ncbi.nlm.nih.gov/COG/) and only eight additional genes were ubiquitous in procaryotes. Another 126 COGs were found in the remaining five microbial genomes, excluding the mycoplasmas, which have a reduced genomic complement. The majority of the universally conserved COGs (65 out of 110) belong to the information storage and processing proteins, which appear to hold more promise for future phylogenetic analysis than the metabolic proteins. However, about half

Editorial Note: Software available from O. Strunk and W. Ludwig, Department of Microbiology, Technische Universität, München, Munich, Germany. ARB is a software environment for sequence data located at: www.mikro.biologie.tu-muenchen.de/pub/ARB.

of these information processing COGs contain ribosomal proteins, which are small and therefore not sufficiently informative for the inference of global phylogenies. This leaves us with about 40–100 genes that fulfill the basic requirements of useful phylogenetic markers.

It has been proposed that many genes involved in the processing of genetic information (components of the transcription and translation systems) exhibit concurrent evolution due to their housekeeping function (Olsen and Woese, 1997). It appears logical that these key systems would be optimized early and then conserved to confer maximum survival and evolutionary benefit on the organism.

Although the databases of alternative phylogenetic markers are small relative to that of the SSU rRNA, some of the other requirements for markers, including representation of phylogenetically diverse organisms, are met by, for example, LSU rRNA, elongation factor Tu/1α, the catalytic subunit of the proton translocating ATPase, *recA*, and the hsp 60 heat shock protein. For some other markers fulfillment of the ubiquity requirement can not be assessed because of the limited state of the sequence databases.

Some Drawbacks of Alternative Phylogenetic Markers

Lateral gene transfer and gene duplication Comparative analyses of the 18 published complete microbial genome sequences does not reveal a consensus picture of the root of the tree of life (Klenk et al., 1997b) or of the relative branching order of the early lineages within the domains. This contradicts the marked separation of the primary domains based on morphology, physiology, biochemical characteristics, and overall genome sequence data. A monophyletic origin of the domain *Archaea* has been put in question by some authors (Gupta, 1998), but genomic evidence for monophyly of this group has also been reported (Gaasterland and Ragan, 1999). This contradiction has led to the assumption that lateral gene transfer and/or gene duplications, often followed by the loss of one or more gene variants in different lineages, has occurred in some potential marker molecules, especially genes coding for proteins involved in central metabolism (Brown and Doolittle, 1997). Obviously, such genes or markers cannot be used for testing major phylogenies deduced from SSU rRNA data.

The usefulness of many proteins as potential phylogenetic markers is curtailed by the presence of duplicated genes in certain organisms. The degree of sequence divergence in these duplicated markers ranges from the interdomain level, as shown for the catalytic subunit of vacuolar and F_1F_0-ATPases of *Enterococcus hirae*, to the species level, exemplified by EF-Tu of *Streptomyces ramocissimus*. When conserved proteins are used as phylogenetic markers for inferring intradomain phylogenies, one has to take care that orthologous genes (common origin) rather than paralogous genes (descendants of duplications) are compared. The recognition of paralogous genes is a central problem in phylogenetic analyses, especially when only limited data sets are available as in the case of the catalytic subunit of the proton-translocating ATPase. Although the sequence similarities between bacterial F_1F_0 type, and archaeal and eucaryal vacuolar type, ATPase subunits are rather low (around 20%), it was initially assumed that the corresponding subunits (β and A or α and B) are homologous molecules (Iwabe et al., 1989; Ludwig et al., 1993). The presence of an F_1F_0 type ATPase β-subunit gene has been shown for all representatives of the domain *Bacteria* investigated thus far (Ludwig et al., 1993; Neumaier, 1996). However, the finding that *Thermus* and other members of the *"Deinococcus-Thermus"* phylum contain vacuolar type ATPases (Tsutsumi et al., 1991; Neumaier, 1996) threatened this ATPase-based phylogenetic picture. It was later found that genes for subunits of vacuolar type ATPases exist in many (but not all) bacterial species from different phyla in addition to the corresponding F_1F_0 type ATPase subunit genes (Kakinuma et al., 1991; Neumaier, 1996). It is commonly accepted that F_1F_0 type ATPase subunits α and β resulted from an early gene duplication and should be regarded as paralogous. The same is assumed for the vacuolar type ATPase subunits A and B. The findings described above suggest additional early gene duplications probably leading to the ancestors of F_1F_0 and vacuolar type ATPase (subunits). Whereas α and β, or A and B subunits, coexist in all cases investigated so far, this is not the case for the F_1F_0 and vacuolar type paralogs. The available data indicate that the former would have become the essential energy-gaining version in the bacterial domain, the latter in the archaeal and eucaryal domains. During the course of evolution, the other member of the duplicate pair apparently changed its function (Kakinuma et al., 1991) and may have lost its essential character. Therefore, the nonessential copy could have been lost by many (even closely related) organisms during the course of evolution. The *"Deinococcus–Thermus"* phylum, in which only vacuolar type ATPases have been found, might be an exception. Assuming an early diversification of the bacterial phyla, the functional diversification of the duplicated ATPases could have occurred during this era of evolution. The ancestor of the members of the *"Deinococcus–Thermus"* phylum may have lost the F_1F_0 version early in evolution. However, early lateral gene transfers as postulated by some authors (Hilario and Gogarten, 1993) cannot be excluded.

There are other examples of gene duplications and premature phylogenetic misinterpretations, as documented by the history of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) based phylogenetic investigations (Martin and Cerff, 1986; Brinkmann et al., 1987; Martin et al., 1993; Henze et al., 1995). Besides these early gene duplications, there are also indications of more recent events, such as the EF-Tu of *Streptomyces*, *hsp60* of *Rhizobium*, or *recA* of *Myxococcus*. Paralogous genes occurring as a result of gene duplication or lateral gene transfer can only be recognized as such in organisms which have preserved more than one version of the (duplicated) gene. And even then it may remain difficult or even impossible to decide which genes can be regarded as orthologous. Obviously, only the orthologous gene, which represents the functionally essential compound, can be used for inferring or evaluating phylogenies. Thus, whenever new potential phylogenetic markers are investigated and major discrepancies with rRNA-based conclusions are found, a comprehensive data base should be established, accompanied by an extensive search for potential gene duplications.

Limited information content Based on currently available sequence data, the LSU rRNA is the only marker which carries more phylogenetic information than the small subunit rRNA. There are more than twice as many informative residues in the large subunit rRNA (Ludwig et al., 1998b). In the case of protein markers, the amino acid sequences are preferred over the coding gene sequences for phylogenetic analysis. Proteins provide the function, and consequently the amino acid sequences are the targets of evolutionary selective pressure. In contrast, the DNA sequence differences, especially at third base positions, are under pressure of the codon preferences of the particular organism. Most of the proteins recognized as useful phylogenetic markers comprise less informative primary structure sites than the rRNA markers. For example, EF-Tu/-1α and ATPase catalytic subunit protein primary structures contain 311 and 359 informative residues, respectively. This deficiency could be partly compensated for by the 20 possible character states (amino acids) per position. However, in real data the number of allowed character states—the positional variability—is reduced due to functional constraints. The current data sets (EF-Tu/-1α and ATPase catalytic subunit) do not contain positions at which more than 15 different amino acids occur, and the largest fractions of positions (18%–20%, 11%–12%, 9%–12%) are represented by positions with only 2, 3, or 4 different residues, respectively.

Conflicting tree topologies Identical tree topologies cannot be expected from phylogenetic analysis of different markers. Given the low phylogenetic information content of each of the markers, and the wide grid of resolution, it is unlikely that independently evolving markers have preserved information on the same eras of evolutionary time. In principle, one would expect that this missing phylogenetic information would yield reduced resolution but not change the topology of the tree. However, the latter is often the case, as shown in Fig. 4. A small stretch of aligned real 16S rRNA sequences was used to generate the tree in Fig. 4A. If it is assumed that this tree illustrates the phylogenetic truth and that the information for the common origin of Spec. 4 and Spec. 5 was lost during the course of evolution, one would expect a reduction in resolution. Removing the alignment column (marked by an arrowhead) responsible for this relationship should result in shortening or deleting of the common branch of Spec. 4 and Spec. 5, producing a multifurcation as shown in Fig. 4B. However, due to branch attraction by residues at other alignment positions the branches of Spec. 4 and Spec. 5 are separated as shown in Fig. 4C, misleadingly simulating a different history. Consequently, local differences of resolution and topology in trees derived from alternative phylogenetic markers do not necessarily indicate a different path of evolution.

Alternative Gene Trees

Large subunit rRNA As alluded to above, the LSU rRNA may be the most informative alternative phylogenetic marker. The primary structure of this molecule is at least as conserved as that of the SSU rRNA, and it contains more and longer stretches of informative positions. The spectrum of the LSU rRNA database is superior to that of all other alternative (protein) markers. Given that both rRNAs are involved in the translation process, it can be assumed that a similar selective pressure has been exerted on both genes. Consequently, LSU rRNA should be more useful for supporting rather than evaluating SSU rRNA-based conclusions. The internal structure (branching orders of the major lineages) of the intradomain trees can also be evaluated, given the availability of representative data sets for both molecules. The overall topologies of trees based upon the sequences of small and large subunit rRNA genes are in good agreement (De Rijk et al., 1995; Ludwig et al., 1998b). A 23S rRNA-based bacterial phyla tree is shown in Fig. 5, the corresponding 16S rRNA-based tree in Fig. 6. Slight local differences between trees reconstructed from both genes with the same method and parameters have been documented (De Rijk et al., 1995; Ludwig et al., 1995). This finding does not really cast doubt on the SSU rRNA-based branching patterns but rather underlines the previously mentioned limitations of phylogenetic markers. The LSU rRNA might be the better phylogenetic marker, providing more information and greater resolution, but the major drawback of this molecule is the currently limited database. Unfortunately, this database has not grown as fast as that for the SSU rRNA.

Elongation factors The elongation factors are also intrinsic components of the translation process but are functionally different from the rRNAs. It is generally assumed that the different classes of elongation (and probably initiation) factors are paralogous molecules resulting from early gene duplications. At present, a reasonable data set is available for EF-Tu/1α. In general, EF-Tu/1α-based domain trees (Fig. 7) globally support rRNA-derived branching patterns (Ludwig et al., 1998b). However, some general problems of protein markers are also exhibited by EF-Tu/1α sequences. As with the rRNA markers, no significant relative branching order for the major intradomain lines of descent can be determined. No major contradictions, e.g., members of a given phylum defined by rRNA sequences clustering among representatives of another phylum, were seen between rRNA and EF-Tu/1α tree topologies. However, in detailed

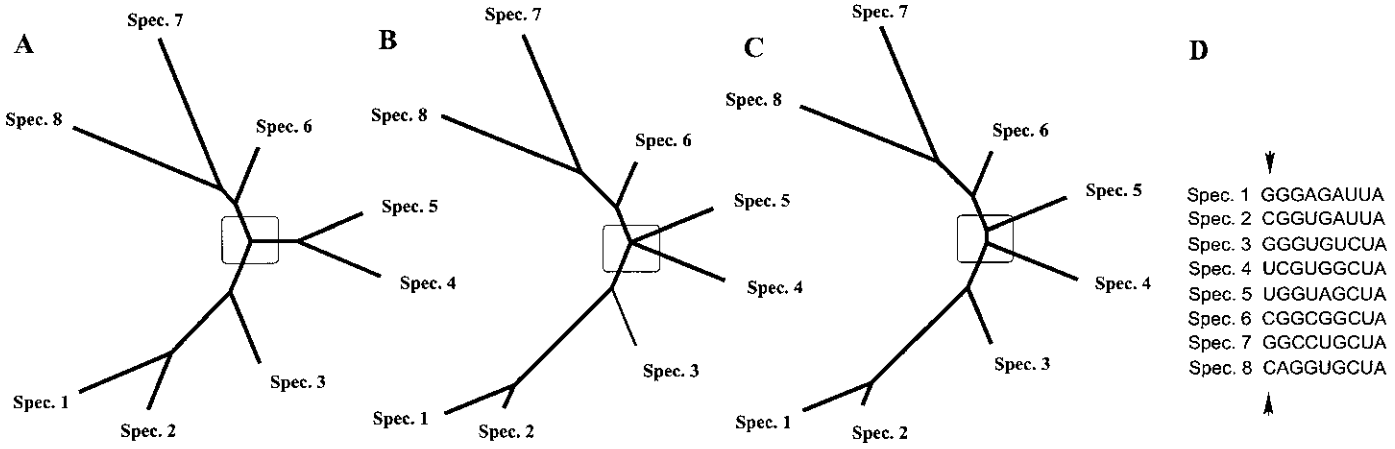

FIGURE 4. Missing phylogenetic information. If the tree in *A* reflects the true phylogeny, a tree topology showing a multifurcation for Spec. 4 and Spec. 5 as well as the other subtrees as shown in *B* would be correct if the phylogenetic information on the monophyletic origin of Spec. 4 and Spec. 5 was not preserved in present day sequences. This can be simulated by exclusion of the column marked by arrowheads in *D*. The loss of this information produces the misleading tree topology of *C* as a result of branch attraction.

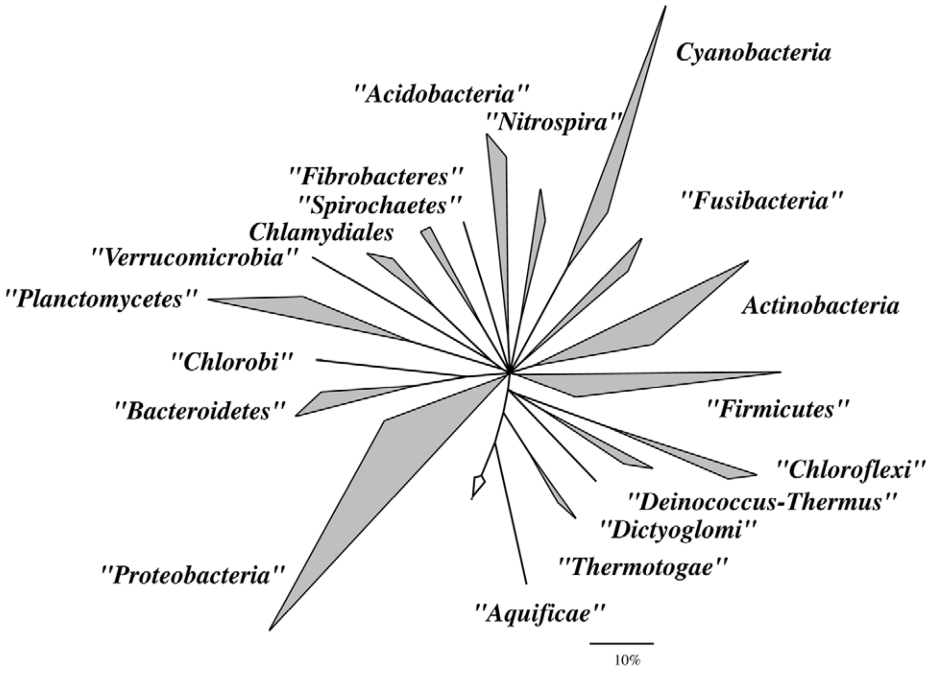

FIGURE 5. 23S rRNA based tree depicting the major bacterial phyla. The triangles indicate groups of related organisms, while the angle at the root of the group roughly indicates the number of sequences available and the edges represent the shortest and longest branch within the group. The tree was reconstructed, evaluated and optimized using the ARB parsimony tool. Only sequence positions sharing identical residues in at least 50% of all bacterial sequences were included in the calculations. All available almost complete homologous sequences from *Archaea* and *Eucarya* were used as outgroup references to root the tree (indicated by the *arrow*). Multifurcations indicate that a relative branching order could not be defined.

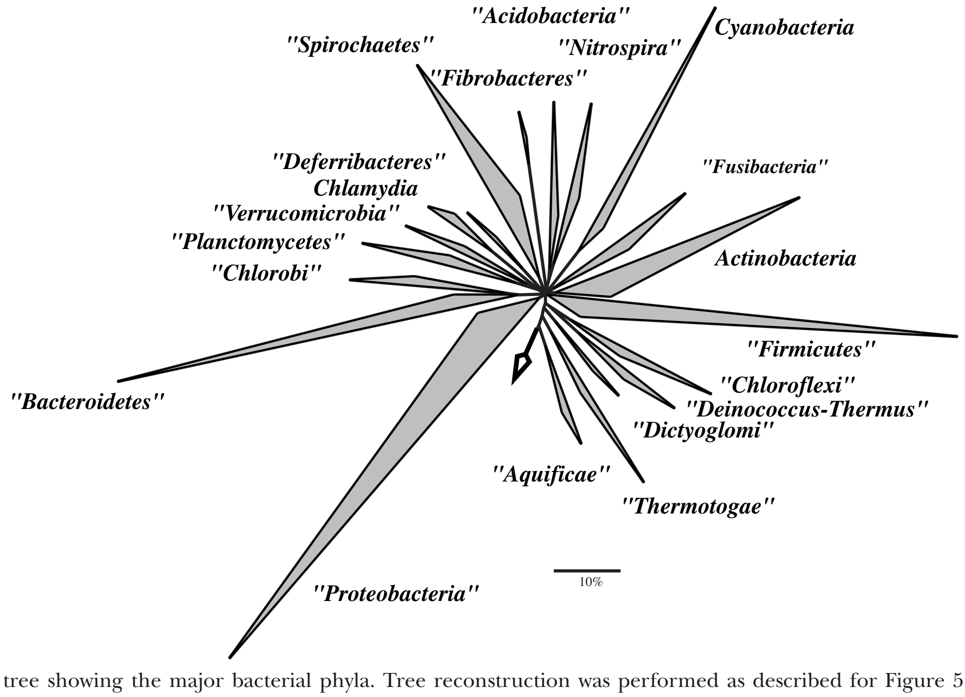

FIGURE 6. 16S rRNA based tree showing the major bacterial phyla. Tree reconstruction was performed as described for Figure 5. Tree layout of this and subsequent trees was according to the description for Figure 5.

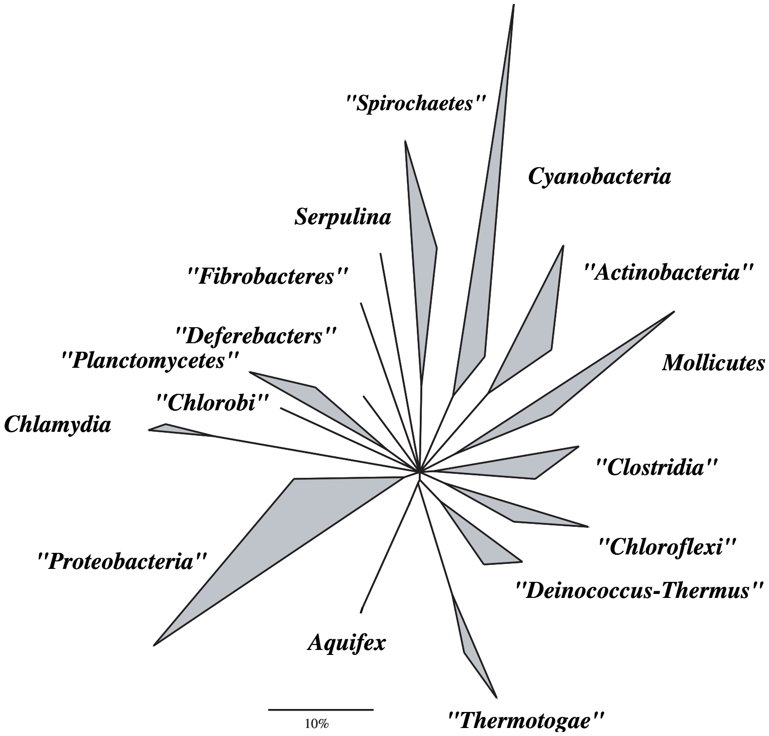

FIGURE 7. Elongation factor Tu based tree illustrating relationships among the major bacterial phyla. The tree was reconstructed from amino acid sequence data, and evaluated and optimized using the ARB parsimony tool. The tree is shown as unrooted, and only positions sharing identical residues in at least 30% of all sequences were included in the calculations.

trees local topological differences have been demonstrated (Ludwig et al., 1993). The reduced phylogenetic information content of EF-Tu (656 bits versus 1506 bits in the SSU rRNA; Ludwig et al., 1998b) may be responsible for the fact that the monophyletic status of some phyla such as the *"Proteobacteria"* is not supported by the protein-based trees. The separation of subgroups such as proteobacterial classes, however, is globally in agreement with the rRNA-based trees.

Interdomain sequence similarities for the rRNAs are 50% and higher, allowing the rooting and (at least to some extent) structuring of the lower branches for a given domain tree versus the other two. The interdomain protein similarities of the elongation factors are low (not more than 30%), making a reliable rooting or structuring of the bacterial tree difficult. The elongation factor database also contains examples of paralogy resulting from gene duplications or lateral gene transfer (Vijgenboom et al., 1994).

RNA polymerases The DNA-directed RNA polymerases (RNAPs) are essential components of the transcription process in all organisms, and the genes for the largest subunits (β and β′ in *Bacteria*; A′, A′′ and B in *"Crenarchaeota"*; B′ and B′′ in *"Euryarchaeota"*) are highly conserved and ubiquitous. The public databases contain RNAP sequences for about 40 species of *Bacteria* and 10 species of *Archaea*. The genes coding for RNAPs are located next to each other on the chromosomes of both *Bacteria* and *Archaea*, and contain 2300 (*Archaea*) to 2400 (*Bacteria*) amino acids that can be clearly aligned for phylogenetic purposes (Klenk et al., 1994). No paralogous genes are known for RNAPs. In general, for the *Bacteria* the intradomain topology of the trees derived from both RNAP large subunits supports the 16S rRNA-based tree in almost all details, with only one major discrepancy: the position of the root of the domain. Intensive rooting experiments with a variety of archaeal and/or eucaryotic outgroups does not place the root of the *Bacteria* close to the extreme thermophiles (*Aquifex* or *Thermotoga* species) as in the rRNA tree, but next to *Mycoplasma* (Klenk et al., 1999). Since the placement of a root within a phylogenetic tree is not critical for most taxonomic purposes, it can be concluded that rRNAs and RNAPs in general support the same intradomain branching pattern for the *Bacteria*.

Proton translocating ATPase The catalytic subunit of proton-translocating ATPase is another example of a protein marker for which a reasonable data set is available, at least with respect to the spectrum of bacterial phyla (Ludwig et al., 1993, 1998b; Ludwig and Schleifer, 1994; Neumaier et al., 1996). This marker should be more appropriate than elongation factors or RNA polymerases for testing the validity of rRNA-based trees for organismal phylogeny, as the ATPase has nothing in common functionally with transcription or translation except its own synthesis.

In general, the F_1F_0 ATPase β-subunit data support the rRNA-based tree (Fig. 8), but the information content and resolving power is reduced. Again, local differences in branching patterns have been shown, and the monophyletic structure of some phyla, defined by rRNA analysis, is not supported (Ludwig et al., 1993).

A correct rooting of the ATPase β-subunit-based bacterial domain tree with the paralogous catalytic subunit of the vacuolar type ATPase (Hilario and Gogarten, 1998; Ludwig et al., 1998b) is not possible as the overall sequence similarities between the two paralogs are not higher than 23%.

There are not sufficient data available for the F_1F_0 ATPase α-subunit (most likely the paralogous pendant of the β-subunit) to allow effective comparison with the rRNA data. However, the currently available α-subunit data set does not indicate great dif-

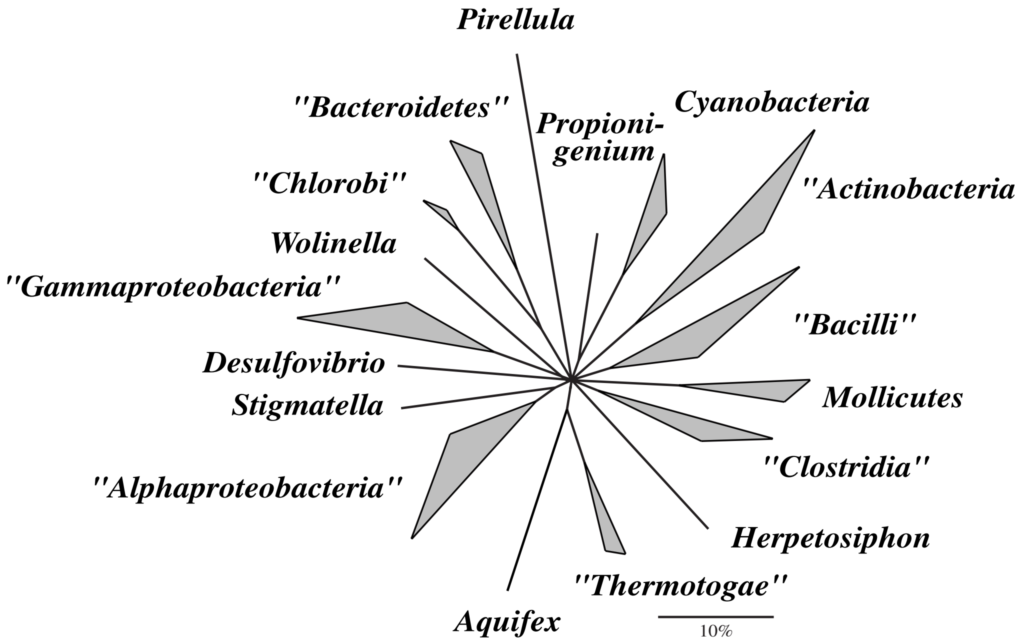

FIGURE 8. F$_1$F$_0$ ATPase β-subunit-based tree depicting the major bacterial phyla. The tree is shown as unrooted. Tree reconstruction was performed as described for Figure 7.

ferences in phylogenetic conclusions inferred from the two data sets. There are also insufficient data for the paralogous subunits A and B of the vacuolar type ATPase; however, a clear separation of the *Eucarya* from the *Bacteria* and *Archaea* is seen when the currently available data set is analyzed. The bacterial and archaeal lines appear intermixed at the lowest level of the corresponding subtree. At present, this intermixing cannot be proven or correctly interpreted (Neumaier, 1996). There is low significance for any branching pattern at this level of (potential) relatedness. Furthermore, only a few positions, which currently cannot be tested for plesiomorphy, are responsible for this intermixing. In addition, functional constancy can not be assumed for eucaryal and archaeal versus bacterial vacuolar type ATPases, and lateral gene transfer cannot be excluded (Hilario and Gogarten, 1993).

recA *protein* Most of the bacterial phyla are represented by one or a few sequences in the *recA* protein sequence data base (Wetmur et al., 1994; Eisen, 1995; Karlin et al., 1995). Comparative analysis of these data again supports the rRNA-based view of bacterial phylogeny. A homologous counterpart for the archaeal and eucaryal phyla has not yet been identified. A significant relative branching order of phyla cannot be defined. Although monophyly of the *"Proteobacteria"* or the Gram-positive bacteria with a low DNA G + C content is not observed, no major contradictions to the rRNA-based phylogeny have been reported. The higher phylogenetic groups (*"Proteobacteria", Cyanobacteria, "Actinobacteria", Chlamydiales, "Spirochaetes", "Deinococcus–Thermus", "Bacteroidetes",* as well as *"Aquificae"*) are separated from each other as in the rRNA-derived phylogeny. However not surprisingly local differences in detailed branching patterns were found.

There is one major discrepancy: phylogenetic analysis of *Acidiphilium* using *recA* sequence data does not show it to cluster within the *"Alphaproteobacteria"* as is found with rRNA analyses. Two *recA* genes, which differ remarkably in sequence, have been found in *Myxococcus xanthus* and may indicate the occurrence of gene duplications or lateral gene transfer. Therefore it is possible that such phenomena have occurred in the evolution of *recA* in *Acidiphilium*.

hsp60 heat shock proteins Sequences for hsp60 chaperonin have been determined for a wide spectrum of bacterial phyla (Viale et al., 1994; Gupta, 1996; 1998). A distant relationship has been postulated for hsp60, the eucaryotic TCP-1 complex, and the archaeal Tf-55 protein (Brown and Doolittle, 1997). However, given the low similarities, the homologous character of hsp60 and the TCP-1 complex or the Tf-55 protein cannot be demonstrated unambiguously.

Trees based upon the currently available hsp60 sequence data set support rRNA-based trees in that the different phyla are well separated from one another, and in cases where several sequences are available for a given phylum, subclusters resemble the rRNA-derived phylogeny. For example, the *"Gammaproteobacteria"* and *"Betaproteobacteria"* are more closely related to one another than to the *"Alphaproteobacteria"* sister group in both hsp60 and rRNA analyses. However, the use of hsp60 as a phylogenetic marker molecule is again complicated by the existence of duplicated genes as, for example, among *Rhizobium* species.

Other supporting and nonsupporting protein markers The hsp70 (70 kDa heat shock protein)-based tree globally supports rRNA-based clustering. The phyla appear to be separated and even the branching order of the classes of the *"Proteobacteria"* (*"Alphaproteobacteria", "Gammaproteobacteria", "Betaproteobacteria"*) is corroborated. The major concern associated with the hsp70-derived phylogeny is the intermixed rooting of bacterial and archaeal major lines of descent (Brown and Doolittle, 1997; Gupta, 1998). No significant branching order can be defined for the intermixed lines, and, as discussed above for the ATPase phylogeny, these findings may reflect missing resolution at the interdomain level.

At first glance, many other proteins (reviewed by Brown and Doolittle, 1997) seem to support the intradomain tree structures of rRNA-based phylogenies. However, meaningful comparative evaluation is difficult due to limitations in phylogenetic information content and/or databases that are insufficient in size and scope. Examples are provided by family B DNA polymerases which might represent useful markers for all three domains, aminoacyl-tRNA synthetases which differ in size and hence in potential information content, and ribosomal proteins which generally are short polypeptides and thus of very limited phylogenetic use (Brown and Doolittle, 1997). Among the enzymes involved in central metabolism, the usefulness of 3-phosphoglycerate kinase is also curtailed by a limited sequence database.

There are a number of potential protein markers for which deduced trees do not clearly support rRNA-based intradomain phylogenetic conclusions, including DNA gyrases and topoisomerases, some enzymes of the central metabolism, and of amino acid synthesis and degradation. However, as no comprehensive sequence databases are available, careful evaluation of the tree topologies is not possible.

Rationale for a 16S rRNA-derived Backbone for *Bergey's Manual*

The introduction of comparative primary structure analysis of the SSU rRNA by Carl Woese and coworkers was undoubtedly a major milestone in the history of systematic biology. This approach opened the door to the elucidation of the evolutionary history of the procaryotes, and provided the first real opportunity to approach the ultimate goal in taxonomy i.e., systematics based upon the natural relationships between organisms. The rapid development of experimental procedures enabled the scientific community to characterize the majority of described species at the 16S rRNA level. During preparation of the new edition of *Bergey's Manual*, coordinated efforts to close the gaps and to investigate the missing species were initiated. There is a realistic prospect of completing the database with respect to all known validly described species in the near future.

Although the resolving power of the SSU rRNA approach has sometimes been overestimated, it has allowed a tremendous expansion in our knowledge of procaryotic relationships during recent years. This has been accompanied by the recognition of limitations in the existing procaryotic taxonomy, and efforts to redress these limitations. The taxonomic history of the pseudomonads is one impressive example of the "phylogenetic cleaning" of a genus that was phylogenetically heterogeneous in composition (Kersters et al., 1996).

It appears that the SSU rRNA is currently the most powerful phylogenetic marker, in terms of information content, depth of taxonomic resolution, and database size and scope. There is also good congruence between global tree topologies derived from different phylogenetic markers, indicating that SSU rRNA-based phylogenetic conclusions indeed reflect organismal evolution, at least at the global level. Local discrepancies in phylogenetic trees resulting from different information content, different rate or mode of change, or inadequate data analysis do not greatly compromise this general picture. The underlying cause of major tree discrepancies may in some cases be the analysis of paralogous genes, as indicated by multiple genes arising from duplication, loss, or lateral transfer of genes.

The logical consequence of these investigations and observations is to structure the present edition of *Bergey's Manual* according to our current (rRNA-based) concept of procaryotic phylogeny, using the global tree topology as a backbone, and to propose an emended framework of hierarchical taxa.

It should be considered that all phylogenetic conclusions and tree topologies presented here are models that represent the present, imperfect view of evolution. The information content of the SSU rRNA database is rather limited for representation of 3–4 billion years of evolution of cellular life. Furthermore, the methods of data analysis and the software and hardware for deciphering and visualizing this information are far from being optimal. For these reasons, the proposed backbone of the taxonomic scheme might be subject to change in the future. The introduction of new taxonomic tools and methods has always had a major impact on contemporaneous taxonomy. New sequence data and improved methods of data analysis may change our view of procaryotic phylogeny. Comparison of previous editions of *Bergey's Manual*, as well as updates of the Approved Lists of Bacterial Names (Skerman et al., 1980), indicates that the contemporary view of microbial taxonomy is determined mainly by the availability, applicability, and resolving power of the methods used to characterize organisms and elucidate their genetic and phylogenetic relationships.

The Small Subunit rRNA-based Tree

The global SSU rRNA-based intradomain phylogenetic relationships are discussed for the *Archaea* and *Bacteria* below. Given that the relationships of these organisms are described in detail in subsequent chapters, only higher phylogenetic levels are shown here. Reconstruction of general trees was performed using only sequences that were at least 90% complete (in relation to the *E. coli* 16S rRNA reference sequence). Lines of descent or phylogenetic groups containing a single or only a few sequences are (usually) not shown in these trees. Environmental sequences from organisms which have not yet been cultured were included in the calculations but are not depicted in the trees. The trees and discussions are based upon a comparative analysis of the current RDP (Maidak et al., 1999) and ARB trees. The RDP tree was reconstructed by applying a maximum likelihood method combined with resampling, whereas for the ARB tree a special maximum parsimony approach in combination with different optimization methods and upper bootstrap limit determination was used. The RDP tree contains the *Bacteria* and *Archaea*, while the ARB tree also includes the *Eucarya*. In both cases, the rooting and internal structuring of the domain trees was estimated using the full data set of the other domains. Although these trees were reconstructed using different methods, their global topologies are in good agreement.

A statistically significant relative branching order cannot be unambiguously determined for the majority of the phyla in the *Bacteria*, or for many of the intraphylum groups, as indicated by multifurcations within the trees. However, clustering tendencies are common to both trees. It should also be considered that most phyla were defined in the early days of comparative rRNA sequencing (Woese, 1987) when the data set was small and long "naked" branches facilitated clear-cut phylum delimitation. With the rapidly expanding database most of these "naked" branches expanded and in some cases it is no longer possible to demonstrate a monophyletic structure or to clearly delimit traditional phyla and other groups, as exemplified by the *"Proteobacteria"* and the a low G + C Gram-positive bacteria (*"Bacilli"*, *"Clostridia"*, *Mollicutes*). The inter- and intra-genus relationships of each group are discussed in detail in subsequent chapters; described below is an overview of the phyla of the bacterial and archaeal domains and their major phylogenetic subclusters (Figs. 9, 10, 11, 12, 13, 14, 15, and 16).

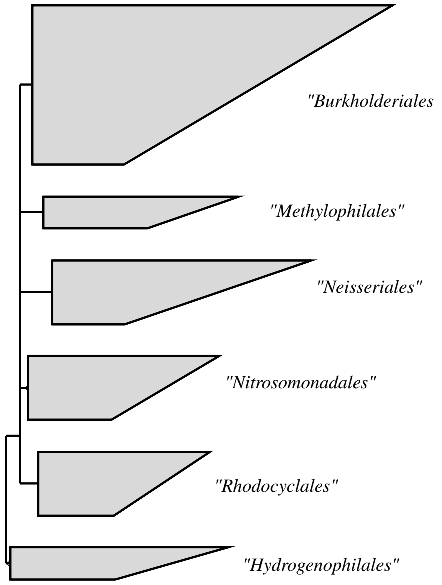

FIGURE 9. 16S rRNA-based tree showing the major phylogenetic groups of the *"Betaproteobacteria"*. Only groups represented by a reasonable number of almost complete sequences are shown. Tree topology is based on the ARB database of 16,000 sequences entries and was reconstructed, evaluated, and optimized using the ARB parsimony tool. A filter defining positions which share identical residues in at least 50% of all included sequences from *"Betaproteobacteria"* was used for reconstructing the tree. The topology was further evaluated by comparison with the current RDP tree, which was generated using a maximum likelihood approach in combination with resampling (Maidak et al., 1999). A relative branching order is shown if supported by both reference trees. Multifurcations indicate that a (statistically) significant relative branching order could not be determined or is not supported by both reference trees.

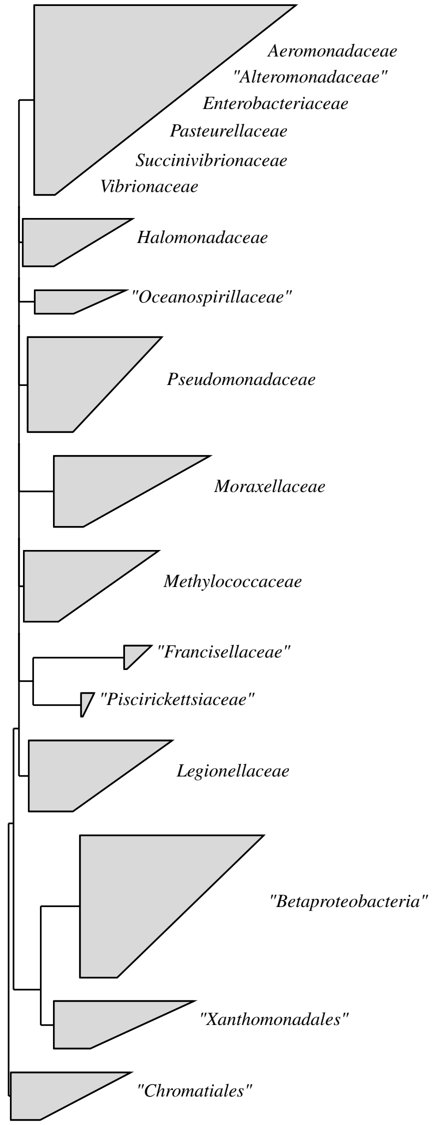

FIGURE 10. 16S rRNA-based tree depicting the major phylogenetic groups within the *"Gammaproteobacteria"*. Tree reconstruction and evaluation was performed as described for Figure 9 with the exception that a 50% filter calculated for the *"Gammaproteobacteria"* was used.

The Bacteria

THE *"Proteobacteria"* The traditional view of the *"Proteobacteria"* as a monophyletic phylum is not completely supported by careful analyses of the current 16S rRNA database. Although there is support for monophyly in the RDP tree, with the *"Deltaproteobacteria"* and *"Epsilonproteobacteria"* forming the deeper branches, a monophyletic structure that includes these two groups is not clearly supported by the ARB tree. Confidence analyses indicate that the significance of a relative branching order within the *"Proteobacteria"* is low in both trees. However, a closer relationship of the *"Gammaproteobacteria"* and *"Betaproteobacteria"*, as well as a common origin of these groups and the *"Alphaproteobacteria"*, is supported by the RDP as well as the ARB tree.

The *"Betaproteobacteria"* (Fig. 9) clearly represents a monophyletic group, comprising the described or proposed higher taxa *"Burkholderiales"*, *"Methylophilales"*, *"Nitrosomonadales"*, *"Neisseriales"*, and *"Rhodocyclales"*. A slightly deeper-branching group comprises the *"Hydrogenophilales"*.

The classical members of the *"Gammaproteobacteria"* (Fig. 10) represent a monophyletic group which includes the *"Betaproteobacteria"* as a major line of descent. In both reference trees the family *"Xanthomonadaceae"* appears to be the most likely sister group of the *"Betaproteobacteria"*. A common clustering of the families *Aeromonadaceae*, *"Alteromonadaceae"*, *Enterobacteriaceae*, *Pasteurellaceae*, and *Vibrionaceae* is supported in both trees. A relative branching order of this cluster and other major groups of the *"Gammaproteobacteria"* such as the families *Halomonadaceae*, *Legionellaceae*, *Methylococcaceae*, *Moraxellaceae*, *"Oceanospirillaceae"*, *Pseudomonadaceae*, and the *"Francisellaceae"*-*"Piscirickettsiaceae"* group cannot be unambiguously determined. In both trees, these groups branch off higher than the *"Betaproteobacteria"*-*"Xanthomonadaceae"* branch, whereas the order *"Chromatiales"* forms a deeper branch. The phylogenetic position of the families *Moraxellaceae* and *Cardiobacteriaceae* relative to that of the *"Gammaproteobacteria"*-*"Xanthomonadaceae"* lineage depends on the treeing method used.

A closer relationship between the families *Rickettsiaceae* and *Ehrlichiaceae* within the *"Alphaproteobacteria"* (Fig. 11) can be seen in both reference trees. The results of tree evaluations indicate branching of this cluster followed by the families *"Sphingomonadaceae"* and the *"Rhodobacteraceae"*. The families *"Bradyrhizobiaceae"*, *Hyphomicrobiaceae*, *"Methylobacteriaceae"*, and *"Methylocystaceae"* represent another subcluster among the *"Alphaproteobacteria"*. A closer interrelated group is formed by the families *Bar-*

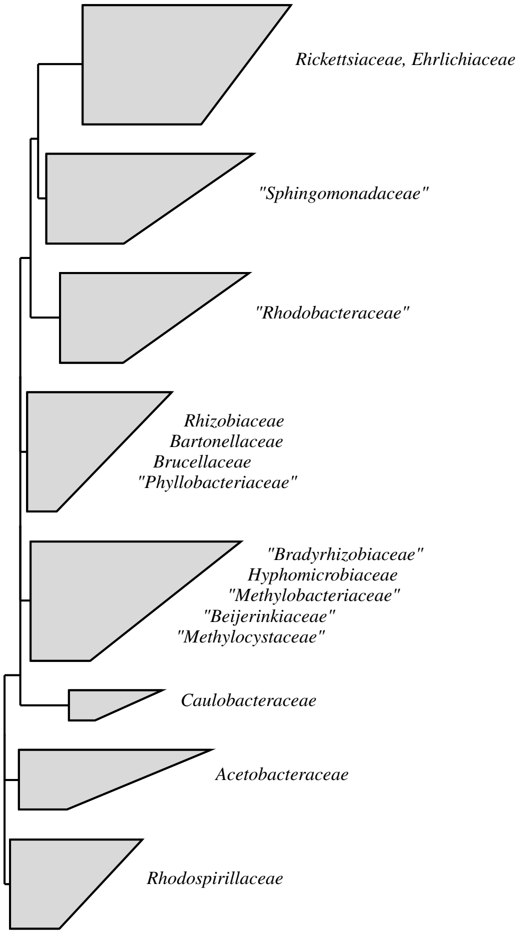

FIGURE 11. 16S rRNA-based tree showing the major phylogenetic groups within the *"Alphaproteobacteria"*. Tree reconstruction and evaluation was carried out as described for Figure 9 with the exception that a 50% filter calculated for the *"Alphaproteobacteria"* was used.

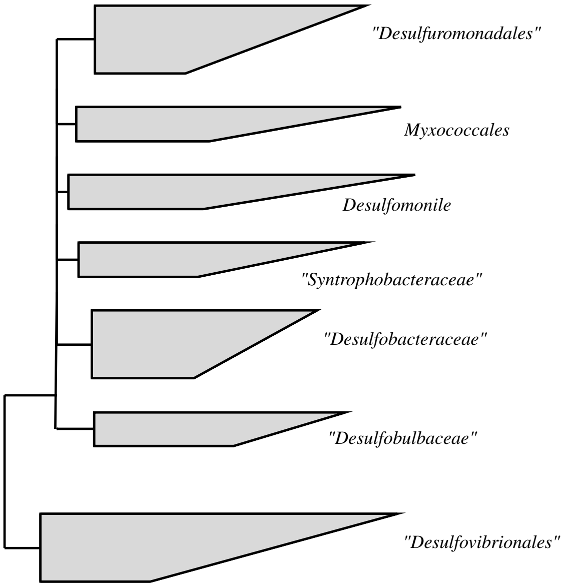

FIGURE 12. 16S rRNA-based tree illustrating the major phylogenetic groups within the *"Deltaproteobacteria"*. Tree reconstruction and evaluation was performed as described for Figure 9 with the exception that a 50% filter calculated for the *"Deltaproteobacteria"* was used.

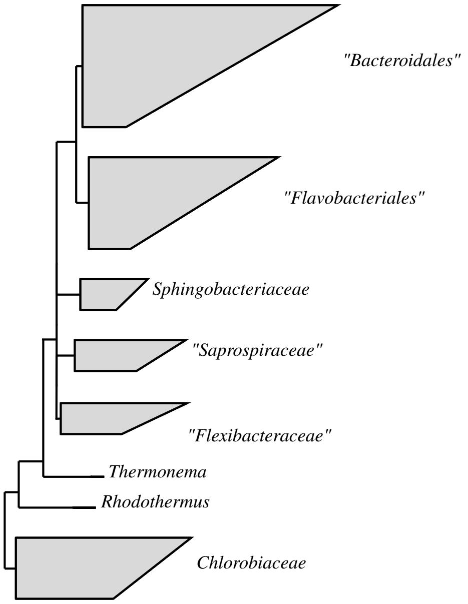

FIGURE 13. 16S rRNA-based tree depicting the major phylogenetic groups within the phylum *"Bacteroidetes"*. Tree reconstruction and evaluation was performed as described for Figure 9 with the exception that a 50% filter calculated for the *"Bacteroidetes"* phylum was used.

tonellaceae, *Brucellaceae*, *Rhizobiaceae*, and *"Phyllobacteriaceae"*. No reliable resolution of these major groups and the family *Caulobacteraceae* can be achieved, but it appears that a deeper branching of the families *Acetobacteraceae* and *Rhodospirillaceae* among the *"Alphaproteobacteria"* is indicated.

The order *"Desulfovibrionales"* currently represents the deepest branch of the *"Deltaproteobacteria"* (Fig. 12). Three other major subgroups comprise *Desulfomonile* and relatives, the *"Syntrophobacteraceae"*, as well as the *"Desulfobulbaceae"*. These subgroups are phylogenetically equivalent in depth to the lineages *"Desulfobacteraceae"*, *"Geobacteraceae"*, and *Myxococcales*.

The families *"Helicobacteraceae"* and *Campylobacteraceae* are the two major lines that form the *"Epsilonproteobacteria"*.

THE *"Spirochaetes"* The *"Spirochaetes"* phylum currently comprises three major subgroups: the sister groups of the families *Spirochaetaceae* and *"Serpulinaceae"*, as well as the deeper branching family *Leptospiraceae*.

"Deferribacteres" AND *"Acidobacteria"* PHYLA To date, the *"Deferribacteres"* phylum is represented by only two cultured species, while only three cultured species are found in the *"Acidobacteria"* phylum. However, a comprehensive data set of environmental sequences indicates a phylogenetic depth and diversity within the *"Acidobacteria"* comparable to that of the *"Proteobacteria"* (Ludwig et al., 1997).

THE *Cyanobacteria* The chloroplast organelles comprise a monophyletic subgroup within the *Cyanobacteria* phylum, which also contains a number of other major lines of descent. The current taxonomy of the cyanobacteria is far from being in accordance with the phylogenetic structure of the phylum.

"Verrucomicrobia", *"Chlamydiae"*, AND *"Planctomycetes"* The phylum *"Verrucomicrobia"* comprises a number of environmental sequences as well as a few cultured members of the genera *Verrucomicrobium* and *Prosthecobacter* (Hedlund et al., 1996). Both reference trees indicate a moderate degree of relationship be-

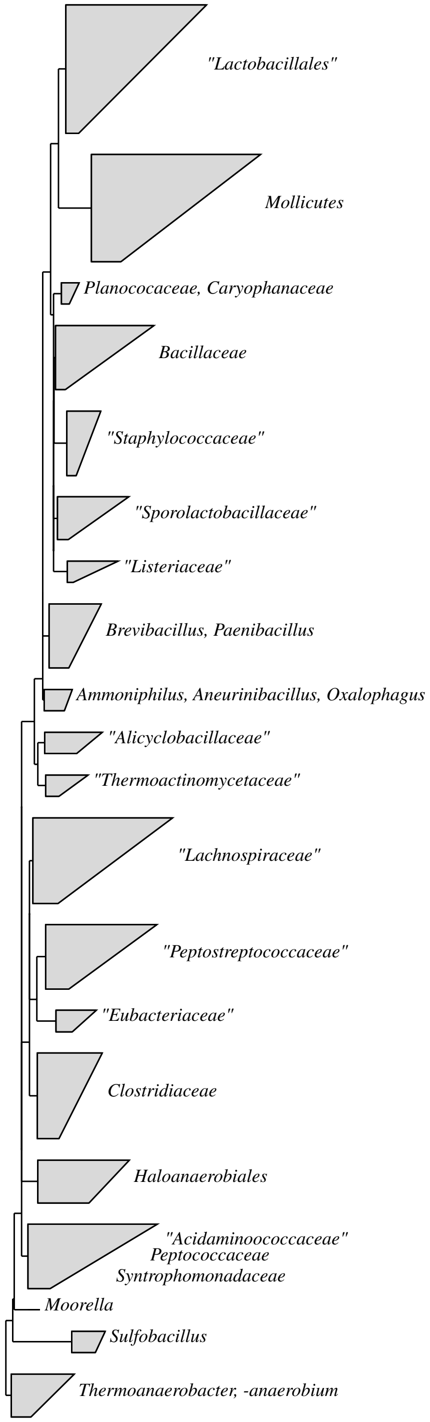

FIGURE 14. 16S rRNA-based tree showing the major phylogenetic groups of the *"Firmicutes"* (Gram-positive bacteria with a low DNA G + C content). Tree reconstruction and evaluation was carried out as described for Figure 9 with the exception that a 50% filter calculated for a core set of sequences (excluding the *Mycoplasmatales* and the deeper groups represented by *Moorella*, *Sulfobacillus*, *Thermoanaerobacter*, and *Thermoanaerobium*) was used.

tween the *"Verrucomicrobia"* and the *Chlamydiales* phylum. However, given the limited number of available sequences for the *"Verrucomicrobia"* and the long naked branch of the *Chlamydiales*, a sister group relationship between these two phyla should be regarded as tentative. A moderate relationship between these two phyla and the *"Planctomycetes"* phylum is also indicated in both the ARB and RDP trees. However, the significance of this branching point is low, and their relationship may not be supported in the future by a growing database. The intraphylum structure of the *"Planctomycetes"* indicates two pairs of sister groups: *Pirellula/Planctomyces* and *Isosphaera/Gemmata*.

"Chlorobi" AND *"Bacteroidetes"* A monophyletic origin of the *"Chlorobi"* (containing the genera *Chlorobium*, *Pelodictyon*, *Prosthecochloris*, and some environmental sequences) and the *"Bacteroidetes"* (Gosink et al., 1998) phyla (Fig. 13) can be seen in both trees and is supported by alternative markers such as large subunit rRNA, and β-subunit of F_1F_0 ATPase. The thermophilic genera *Rhodothermus* and *Thermonema* represent the deepest branches of the phylum *"Bacteroidetes"*. A common root of the *"Bacteroidales"* and *"Flavobacteriales"* within the phylum is supported in both reference trees. This cluster seems to be phylogenetically equivalent to the other major groups i.e., the *Sphingobacteriaceae*, *"Saprospiraceae"*, *"Flexibacteraceae"*, *Flexithrix*, and *Hymenobacter*.

LOW G + C GRAM-POSITIVE BACTERIA Other than for the *"Proteobacteria"*, the most comprehensive 16S rRNA gene sequence database (with more than 1750 almost complete sequences) is available for the Gram-positive bacteria with a low DNA G + C content (*"Bacilli"*, *"Clostridia"*, *Mollicutes*). The common origin of the organisms classically assigned to this group is not significantly supported by all reference trees (see Fig. 14). The *Mollicutes*, comprising the families *Mycoplasmataceae*, *Acholeplasmataceae*, and their walled relatives, represent a monophyletic unit. The classical lactic acid bacteria are members of the families *"Aerococcaceae"*, *"Carnobacteriaceae"*, *"Enterococcaceae"*, *Lactobacillaceae*, *"Leuconostocaceae"*, and *Streptococcaceae*, and are unified in the order *"Lactobacillales"*. A clear resolution of the relationships between the families *Bacillaceae*, *Planococcaceae*, *"Staphylococcaceae"*, *"Sporolactobacillaceae"*, and *"Listeriaceae"* cannot be achieved. Two slightly deeper branching clusters comprise the genera groups of *Brevibacillus*–*Paenibacillus* and *Ammoniphilus*–*Aneurinibacillus*–*Oxalophagus*. The *"Alicyclobacillaceae"* and *Thermoactinomyces* groups represent a further deeper branch. Another major subbranch unifies the *"Eubacteriaceae"*, *Clostridiaceae*, *"Lachnospiraceae"*, and *"Peptostreptococcaceae"* . The *"Eubacteriaceae"* and *"Peptostreptococcaceae"* appear to be sister groups. The phylogenetic position of the order *Haloanaerobiales* is strongly influenced by the treeing method applied and should be regarded as tentative. The families *Haloanaerobiaceae* and *Halobacteroidaceae* constitute a well-defined phylogenetic unit in both reference trees. However, the assignment of this unit to the low G + C Gram-positive phylum is not clearly supported when different treeing methods are applied, suggesting that this group may represent its own phylum. A deeper rooting within the phylum is indicated for the *Peptococcaceae*–*Syntrophomonadaceae* cluster but the phylogenetic position of the genera *Moorella*, *Sulfobacillus*, *Thermoanaerobacter*, and *Thermoanaerobium* is uncertain. The latter two genera represent a phylogenetic unit, but this unit and each of the other genera probably represent additional phyla.

"Fusobacteria" PHYLUM The *"Fusobacteriaceae"* phylum so far comprises only three subclusters: *Fusobacterium*, *Propionigenium*–*Ilyobacter* and *Leptotrichia*–*Sebaldella*.

HIGH G + C GRAM POSITIVE BACTERIA (*"Actinobacteria"*) The phylum of the Gram-positive bacteria with a high G + C DNA content (the *"Actinobacteria"*) provides an example of a clearly defined and delimited major bacterial line of descent. As seen in Fig. 15, the families *Rubrobacteraceae* and *Coriobacteriaceae*

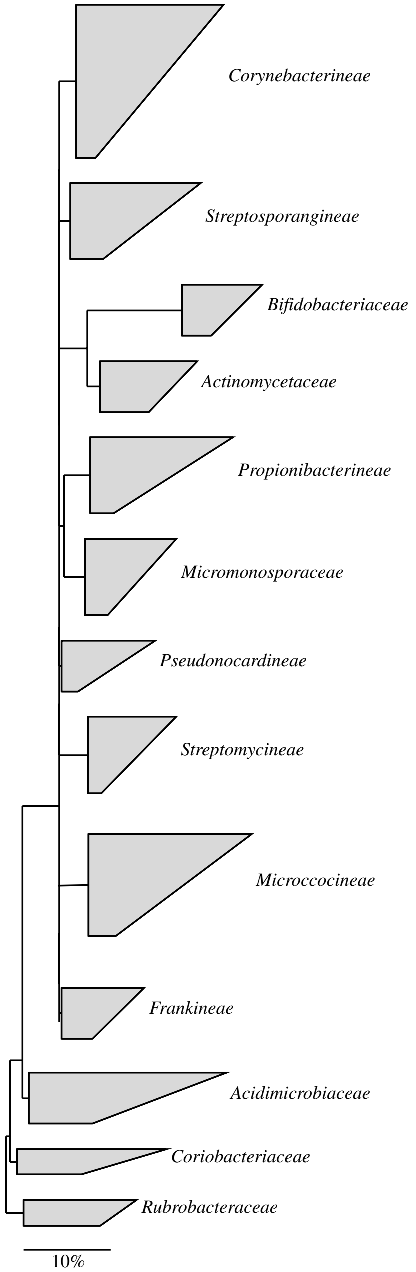

FIGURE 15. 16S rRNA-based tree depicting the major phylogenetic groups within the *"Actinobacteria"* (Gram-positive with a high DNA mol% G + C content). Tree reconstruction and evaluation was performed as described for Figure 9 with the exception that a 50% filter calculated for this phylum was used.

currently represent the deepest branches of the phylum, whereas the family *Acidimicrobiaceae* occupies an intermediate position between the former two and the remaining major subgroups of the phylum. There is some support for a common origin of the *Bifidobacteriaceae* and *Actinomycetaceae*, and for the clustering of the families *Propionibacteriaceae* and *Micromonosporaceae*. No significant or stable branching order for these and other subgroups such as *Corynebacteriaceae*, *Frankineae*, *Pseudonocardiaceae*, *Streptomycetaceae*, and *Streptosporangineae* could be achieved.

Other phyla The *"Nitrospira"* phylum contains a limited number of organisms, namely representatives of the genera *Nitrospira*, *Leptospirillum*, *Thermodesulfovibrio*, and *Magnetobacterium*. Similarly, only a limited number of organisms and environmental sequences represent the phylum of the green non-sulfur bacteria, which includes the families *"Chloroflexaceae"*, *"Herpetosiphonaceae"*, and *"Thermomicrobiaceae"*. Two major subgroups, the *Deinococcaceae* and the *"Thermaceae"*, have been identified within the *"Deinococcus–Thermus"* phylum. The orders *"Thermotogales"* and *"Aquificales"* constitute two of the deeper branching phyla within the bacterial domain.

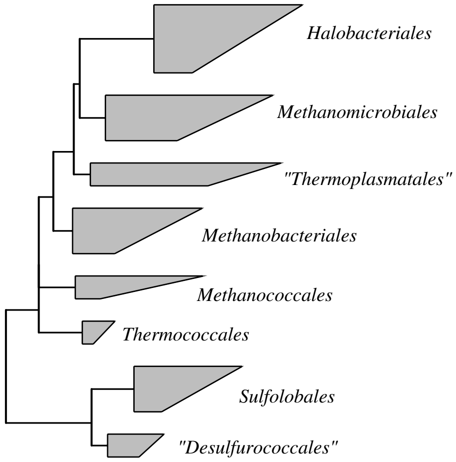

FIGURE 16. 16S rRNA-based tree showing the major phylogenetic groups within the *Archaea*. Tree reconstruction and evaluation was carried out as described for Figure 9 with the exception that tree optimization was performed independently for the *"Euryarchaeota"* and *"Crenarchaeota"* using a 50% filter in each case.

The existence of additional phyla is suggested by the phylogenetic position of organisms such as *Dictyoglomus thermophilum* and *Desulfobacterium thermolithotrophum*, and of some environmental sequences. However, the phylum status of these lineages cannot be evaluated at this time, due to the paucity of available sequence data.

***The* Archaea** Two major lines of descent (phyla) have been delineated within the *Archaea*: the *"Euryarchaeota"*, and the *"Crenarchaeota"*. Within the *"Euryarchaeota"*, the orders *Halobacteriales*, *Methanomicrobiales*, and *"Thermoplasmatales"* share a common root. A relationship between the first two orders is suggested in both reference trees (Fig. 16), and the order *Methanobacteriales* is indicated as the next deepest branch. A stable and significant tree topology resolving the relationship between these four orders and the orders *"Archaeoglobales"*, *Methanococcales*, *Thermococcales*, and *"Methanopyrales"* cannot be deduced from the current database.

The orders *Sulfolobales* and *"Desulfurococcales"* appear to be sister groups within the *"Crenarchaeota"*, while a monophyletic structure of the *Thermoproteales* is somewhat questionable. The genus *Thermophilum* tends to root outside the *Thermoproteales* group, however the significance of this branching is low and the database does not contain sufficient entries to allow careful evaluation of this outcome.

A third archaeal phylum, *"Korarchaeota"*, has been postulated on the basis of two partial environmental 16S rRNA sequences, but representatives of this lineage have not yet been isolated in pure culture. Consequently, the phylum status as well as phylogenetic position of the lineage can currently not be assessed.