Bacterial phylogeny based on 16S and 23S rRNA sequence analysis

W. Ludwig * and K.H. Schleifer

Lehrstuhl für Mikrobiologie der Technischen Universität München, Arcisstr. 21, D-80290 München, FRG

Abstract: Molecular phylogeny increasingly supports the understanding of organismal relationships and provides the basis for the classification of microorganisms according to their natural affiliations. Comparative sequence analysis of ribosomal RNAs or the corresponding genes currently is the most widely used approach for the reconstruction of microbial phylogeny. The highly and less conserved primary and higher order structure elements of rRNAs document the history of microbial evolution and are informative for definite phylogenetic levels. An optimal alignment of the primary structures and a careful data selection are prerequisites for reliable phylogenetic conclusions. rRNA based phylogenetic trees can be reconstructed and the significance of their topologies evaluated by applying distance, maximum parsimony and maximum likelihood methods of phylogeny inference in comparison, and by fortuitous or directed resampling of the data set. Phylogenetic trees based on almost equivalent data sets of bacterial 23S and 16S rRNAs are in good agreement and their overall topologies are supported by alternative phylogenetic markers such as elongation factors and ATPase subunits. Besides their phylogenetic information content, the differently conserved primary structure regions of rRNAs provide target sites for specific hybridization probes which have been proven to be powerful tools for the identification of microbes on the basis of their phylogenetic relationships.

Key words: 16S rRNA; 23S rRNA; Identification; Phylogeny; Systematics

Introduction

Bacterial systematics may be defined as the scientific study of organismal diversity and interrelationships with the ultimate object of characterizing and arranging them in an orderly manner [1]. Taxonomy is often used as a synonym for systematics and consists of classification, nomenclature, and identification [2]. The orderly arrangement of defined taxonomic units (e.g. species) into groups (e.g. genus) is called classification [3]. Nomenclature in bacterial taxonomy means the labelling of units defined by classification according to the International Code of Nomenclature of Bacteria [4]. This code contains rules about the rank of taxa, naming of taxa, nomenclatural types and their designation, priority and publication of names, citation of authors and names, changes in names of taxa as a result of transference, union, or change in rank, rules about illegitimate names and epithets, replacement, rejection, and conservation of names and epithets [1]. An array of different methods is used for the characterization of an organism, whereby a comprehensive collection of data describing its properties is obtained [1]. These data sets may comprise morphological characteristics, physiological properties, chemical composition of cell
constituents such as cell wall, membranes or genomic DNA G + C content, profiles of cellular proteins or genomic restriction fragments, and molecular sequence data. The comparison of organismal characteristics is used to arrange the organisms in groups sharing common properties. The determination of at least part of the most characteristic properties is necessary for the reliable identification of an organism. Classification and identification are often confused with each other. However, it is necessary to describe and characterize first the basic taxonomic unit before an isolate can be identified as belonging to this taxon. On the other hand, when the isolate cannot be identified as a member of a known taxonomic unit, it has to be described and classified as a new taxon.

Since every organism is the product of an evolutionary progression, one has to know its evolutionary history to understand it and to define it in biological terms [5]. It had always been one of the major goals of systematics to establish a classification which reflects the genealogical relationships of the organisms.

Whereas the Linnean classification system could be transformed into a phylogenetic systematics for metazoa, it was not possible to find a basis of phylogenetic relationships for bacterial classification until the 1970s. The morphological complexity of animals and plants, the ontogenic recapitulation of the phylogeny, and the existing knowledge of fossil records allowed to establish, albeit somewhat imperfect, phylogenetically based classification. However, bacterial morphology, physiology and many other properties are not informative enough or too unstable to be used as phylogenetic markers. Furthermore, bacteria do not have meaningful developmental stages, and phylogenetically informative bacterial fossils are not available. Thus, bacterial classification remained a determinative one, despite the efforts of microbiologists to infer a natural bacterial classification [6,7].

The idea which entailed the 'molecular revolution' [8] in bacterial systematics was first introduced in Zuckerkandl and Pauling’s article “Molecules as documents of evolutionary history” [9]. Molecular sequences are documents of evolutionary history. They are essentially linear, and contain hundreds to thousands of more or less independent characters [8]. Since there is an enormous number of possible sequences of a given length or composition, sequence similarity indicates a common origin of the corresponding molecules or genes. The number and nature of sequence differences among functionally corresponding molecules in different organisms reflect the course of evolution and may be used for the reconstruction of molecular genealogies.

Following Zuckerkandl and Pauling’s publication it took a further decade before the molecular approach had been successfully and extensively applied for the elucidation of bacterial phylogeny. The breakthrough in the study of the phylogeny of prokaryotes was achieved by Carl Woese and co-workers in the seventies [8,10,11]. They introduced rapid methods of comparative 16S rRNA sequence analysis and phylogenetic tree reconstruction. The results of these efforts provided, for the first time, insight into the phylogeny of prokaryotes and are a prerequisite for establishing a phylogenetically based classification.

**Nucleic and amino acid sequences as molecular chronometers**

Contemporary organisms are the products of historical events during the course of evolution and their structures at all levels reflect their evolutionary history [8]. In the case of microorganisms, historical documents of at least earlier evolutionary events cannot be found at the phenotypic level. It has taken almost 100 years to finally come close to unravelling phylogenetic relationships of prokaryotes [12]. Comparative analysis of phylogenetic marker molecules now allows the reconstruction of bacterial phylogenies. In comparison with the phenotypic level of evolution, there is a much higher number of definable characters at the molecular and genotypic level. The underlying genetic information for every definable phenotypic function comprises tens to thousands of evolutionary independent sequence residues. Many of these residues can be changed in a neutral fashion without changing or abolis-
ing the overlying phenotypic function. Another advantage of the molecular character states (nucleotides and amino acids) over phenotypic characters is that they are precisely defined. Furthermore, the homologous character of molecules and their functions can often hardly be recognized without molecular information. At the molecular level, sequence similarity indicates the common origin of the corresponding genes and gene products. Number and nature of sequence differences among contemporary molecules or genes reflect the evolutionary history of them. An ideal molecular chronometer would accumulate changes more or less randomly in time. Consequently the number of changes can be used to measure the time that elapsed since the most recent common ancestor passed away. If these changes are selectively neutral and independent of the overlaying phenotype, the ‘evolutionary clock’ [13] embedded in the genotype gives the biologist the capacity to infer evolutionary histories and relationships [14].

Not all genes or macromolecules are useful phylogenetic markers, and not all phylogenetic marker molecules are useful for the analysis of a given group of bacteria or the phylogenetic level of interest. The most reliable molecular chronometers might seem to be randomly evolving ‘functionless’ parts of the genome. However, changes become fixed very rapidly, and since there is only a limited number of possible character states, multiple changes at particular sequence positions cannot be detected, since they may occur frequently with progressing time. Multiple changes definitely obscure the history of the molecules by simulating ‘false’ identities and masking the true number of evolutionary events. Therefore, the space of evolutionary time or the range of organisms over which these sequences give useful information is extremely narrow [13,14]. The most useful molecular chronometers for a given phylogenetic level are molecules which are universal, constant, and highly constraint in function within all organisms related below this level. Depending on their importance for the function of the molecule, individual sequence positions differ in rate and nature of substitution during the course of evolution. Therefore, useful phylogenetic markers usually are ‘nonlinear’ chronometers and phylogenetic conclusions are more problematic than with randomly evolving sequences. Due to this nonlinearity, a direct correlation between the degree of sequence dissimilarity and the elapsed time cannot be postulated. The advantage of the different rates at which positions change is that there are rather conserved to highly variable positions or regions within one molecule which are informative for different phylogenetic levels, i.e. for different eras of evolution. The more conserved regions report on earlier events of evolution, whereas more variable positions carry the information on lower levels of phylogenetic relationships. Presently, the most useful and most extensively investigated phylogenetic marker molecules are the ribosomal RNAs, especially 16S rRNAs and to lesser extent 23S rRNAs.

Size is an important factor regarding the usefulness of molecules for the reconstruction of major phylogenies. Since every individual sequence position can only carry the information on a rather narrow range of evolutionary time, an increasing number of independently evolving positions or regions augments the number of phylogenetic levels which probably can be detected. Another advantage of larger size molecules such as 16S and 23S rRNA is that local non-random rearrangements [15,16] do not disturb the overall picture derived from comparative analysis to such an extent as it has been shown for a small molecule like 5S rRNA [5,17].

**Gene versus whole organism phylogeny**

More than 15 years of comparative 16S rRNA analysis resulted in a large, rapidly growing sequence data set. Presently about 1500 more or less complete 16S rRNA sequences from Bacteria are available to the public at the RDP [18] and Antwerpen databases [19]. These data allow a comprehensive insight in bacterial evolution which had and has impact on the restructuring of bacterial classification. Since most phylogenetic investigations of bacteria have been done on 16S rRNA level, the question arises whether these
data really reflect the organismal phylogeny or only the history of the molecule or gene. The results of comparative analyses of other conserved macromolecules such as elongation factors [20–27], subunits of ATPase [21,28–32], and RNA polymerases [33] are in good agreement with the conclusions drawn from the rRNA data. There are some differences in the topologies of published phylogenetic trees derived from the different marker molecules, but a careful re-evaluation of the data showed that there is less phylogenetic information within the protein molecules than within rRNAs, and that the statistical significance of the differences is low [5]. The resolution power of the protein molecules is not as good as that of the rRNA molecules, particularly for the remotely related organisms. Since similar results have been obtained by analyzing functionally quite different molecules such as rRNAs and ATPases, it is unlikely that rRNA-based phylogenies are artifacts resulting from lateral gene transfer. This does not hold true for non-essential, usually less conserved genes and gene products which may have been subjected to effective lateral gene transfer.

**The rRNA molecule**

**Conservation profile**

The larger rRNA molecules (16S and 23S rRNA) are useful as a measure of phylogenetic relationships, since they provide all prerequisites of phylogenetic marker molecules, i.e. universal distribution, structural and functional conservation, and sufficient size. Their primary structures are alternating sequences of invariant, more or less conserved to highly variable regions. The frequencies of compositional changes at different positions in the molecules vary greatly [34,35]. This fact permits investigation of a wide spectrum of phylogenetic distances from the domain to the species level. Invariant and highly conserved posi-

![Conservation profile of a 23S rRNA region homologous to positions 1521–1542 of *Escherichia coli* 23S rRNA. The profile is based on an alignment of all available complete 23S rRNA sequences from Bacteria. The most frequent character was determined for each individual alignment position. The heights of the graphs indicate the fraction of sequences (%) in which particular characters are present in the four phylogenetically different groups.](image)
Table 1

EMBL database accession numbers, phylogenetic affiliation, length and base composition (%) of bacterial 23S rRNAs

<table>
<thead>
<tr>
<th>Organism</th>
<th>Accession No.</th>
<th>Phylum</th>
<th>Length</th>
<th>G + C</th>
<th>G + U</th>
<th>G + A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermotoga maritima</td>
<td>M67498</td>
<td>Ttg</td>
<td>3029</td>
<td>69.2</td>
<td>51.3</td>
<td>58.0</td>
</tr>
<tr>
<td>Thermus thermophilus</td>
<td>X13612</td>
<td>Dei</td>
<td>2914</td>
<td>63.8</td>
<td>50.9</td>
<td>58.4</td>
</tr>
<tr>
<td>Pelobatella marina</td>
<td>X07408</td>
<td>Pla</td>
<td>2877</td>
<td>52.7</td>
<td>52.0</td>
<td>57.0</td>
</tr>
<tr>
<td>Borrelia burgdorferi</td>
<td>M93664</td>
<td>Spi</td>
<td>2926</td>
<td>44.3</td>
<td>52.1</td>
<td>58.6</td>
</tr>
<tr>
<td>Leptospira interrogans</td>
<td>X14249</td>
<td>Spi</td>
<td>2958</td>
<td>51.3</td>
<td>51.3</td>
<td>57.7</td>
</tr>
<tr>
<td>Anaerococcus laidani</td>
<td>X00512</td>
<td>Cya</td>
<td>2876</td>
<td>54.2</td>
<td>50.6</td>
<td>58.6</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>V00351</td>
<td>Pga</td>
<td>2904</td>
<td>53.4</td>
<td>51.8</td>
<td>57.6</td>
</tr>
<tr>
<td>Pseudomonas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>perenitomarina</td>
<td>L03788</td>
<td>Pga</td>
<td>2894</td>
<td>51.7</td>
<td>52.1</td>
<td>57.0</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>Y00432</td>
<td>Pga</td>
<td>2893</td>
<td>52.7</td>
<td>52.4</td>
<td>57.3</td>
</tr>
<tr>
<td>Ruminobacter amyphilus</td>
<td>X06765</td>
<td>Pga</td>
<td>2867</td>
<td>49.6</td>
<td>51.7</td>
<td>59.9</td>
</tr>
<tr>
<td>Neisseria gonorrhoeae</td>
<td>X67293</td>
<td>Pbe</td>
<td>2890</td>
<td>52.2</td>
<td>51.4</td>
<td>57.7</td>
</tr>
<tr>
<td>Neisseria meningitidis</td>
<td>X67300</td>
<td>Pbe</td>
<td>2890</td>
<td>51.7</td>
<td>51.2</td>
<td>57.7</td>
</tr>
<tr>
<td>Pseudomonas cepacia</td>
<td>X16368</td>
<td>Pbe</td>
<td>2878</td>
<td>53.0</td>
<td>52.4</td>
<td>57.3</td>
</tr>
<tr>
<td>Bordetella actin</td>
<td>X70370</td>
<td>Pbe</td>
<td>2865</td>
<td>52.7</td>
<td>52.0</td>
<td>57.2</td>
</tr>
<tr>
<td>Bordetella bronchiseptica</td>
<td>X70371</td>
<td>Pbe</td>
<td>2878</td>
<td>53.0</td>
<td>52.3</td>
<td>57.6</td>
</tr>
<tr>
<td>Bordetella pertussis</td>
<td>X68323</td>
<td>Pbe</td>
<td>2874</td>
<td>53.0</td>
<td>51.8</td>
<td>57.7</td>
</tr>
<tr>
<td>Bordetella parapertussis</td>
<td>X68368</td>
<td>Pbe</td>
<td>2875</td>
<td>53.0</td>
<td>51.7</td>
<td>57.8</td>
</tr>
<tr>
<td>Rhodobacter capsulatus</td>
<td>X06485</td>
<td>Pal</td>
<td>2742</td>
<td>54.2</td>
<td>50.0</td>
<td>57.0</td>
</tr>
<tr>
<td>Rhodobacter sphaeroides</td>
<td>X53853</td>
<td>Pal</td>
<td>2742</td>
<td>55.9</td>
<td>49.9</td>
<td>57.4</td>
</tr>
<tr>
<td>Mycobacterium leprae</td>
<td>X56657</td>
<td>Gph</td>
<td>3132</td>
<td>55.1</td>
<td>52.3</td>
<td>56.8</td>
</tr>
<tr>
<td>Mycobacterium kansasi</td>
<td>Z12712</td>
<td>Gph</td>
<td>3123</td>
<td>56.7</td>
<td>51.4</td>
<td>57.3</td>
</tr>
<tr>
<td>Micrococcus luteus</td>
<td>X06484</td>
<td>Gph</td>
<td>3093</td>
<td>55.1</td>
<td>53.5</td>
<td>57.1</td>
</tr>
<tr>
<td>Frankia sp.</td>
<td>M55343</td>
<td>Gph</td>
<td>3108</td>
<td>57.4</td>
<td>54.5</td>
<td>57.1</td>
</tr>
<tr>
<td>Streptomyces griseus</td>
<td>M76388</td>
<td>Gph</td>
<td>3128</td>
<td>56.4</td>
<td>54.0</td>
<td>56.9</td>
</tr>
<tr>
<td>Streptomyces antibioticus</td>
<td>M27245</td>
<td>Gph</td>
<td>3120</td>
<td>57.0</td>
<td>54.1</td>
<td>56.9</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>K006317</td>
<td>Gpl</td>
<td>2928</td>
<td>53.9</td>
<td>50.9</td>
<td>57.9</td>
</tr>
<tr>
<td>Bacillus globisporus</td>
<td>X06824</td>
<td>Gpl</td>
<td>2932</td>
<td>52.8</td>
<td>51.0</td>
<td>57.6</td>
</tr>
<tr>
<td>Bacillus licheniformis</td>
<td>X06833</td>
<td>Gpl</td>
<td>2932</td>
<td>54.7</td>
<td>50.8</td>
<td>58.0</td>
</tr>
<tr>
<td>Bacillus anthracis</td>
<td>X63645</td>
<td>Gpl</td>
<td>2904</td>
<td>51.6</td>
<td>50.0</td>
<td>57.4</td>
</tr>
<tr>
<td>Bacillus sp.</td>
<td>X60981</td>
<td>Gpl</td>
<td>2927</td>
<td>57.8</td>
<td>51.2</td>
<td>59.0</td>
</tr>
<tr>
<td>Bacillus stetheromophilus</td>
<td>X01387</td>
<td>Gpl</td>
<td>2928</td>
<td>59.6</td>
<td>50.9</td>
<td>58.8</td>
</tr>
<tr>
<td>Streptococcus thermophilis</td>
<td>X06829</td>
<td>Gpl</td>
<td>2901</td>
<td>49.8</td>
<td>51.7</td>
<td>57.8</td>
</tr>
<tr>
<td>Streptococcus oralis</td>
<td>X06827</td>
<td>Gpl</td>
<td>2902</td>
<td>49.8</td>
<td>51.5</td>
<td>57.0</td>
</tr>
<tr>
<td>Lactococcus lactis</td>
<td>X66335</td>
<td>Gpl</td>
<td>2901</td>
<td>49.7</td>
<td>51.4</td>
<td>57.9</td>
</tr>
<tr>
<td>Lactococcus delbrueckii</td>
<td>X66342</td>
<td>Gpl</td>
<td>2909</td>
<td>52.8</td>
<td>50.5</td>
<td>59.0</td>
</tr>
<tr>
<td>Staphylococcus carnosus</td>
<td>X06841</td>
<td>Gpl</td>
<td>2924</td>
<td>50.8</td>
<td>52.0</td>
<td>57.3</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>X06842</td>
<td>Gpl</td>
<td>2923</td>
<td>50.0</td>
<td>51.9</td>
<td>57.2</td>
</tr>
<tr>
<td>Listeria monocytogenes</td>
<td>X06840</td>
<td>Gpl</td>
<td>2932</td>
<td>52.0</td>
<td>50.8</td>
<td>58.6</td>
</tr>
<tr>
<td>Clostridium tyrobutyricum</td>
<td>X06842</td>
<td>Gpl</td>
<td>2886</td>
<td>50.6</td>
<td>50.4</td>
<td>58.3</td>
</tr>
<tr>
<td>Clostridium botulinum</td>
<td>X065002</td>
<td>Gpl</td>
<td>2895</td>
<td>50.7</td>
<td>50.5</td>
<td>57.7</td>
</tr>
<tr>
<td>Pectinatus frisingensis</td>
<td>X06423</td>
<td>Gpl</td>
<td>2924</td>
<td>52.7</td>
<td>48.4</td>
<td>60.9</td>
</tr>
<tr>
<td>Peptococcus niger</td>
<td>X06428</td>
<td>Gpl</td>
<td>2900</td>
<td>51.8</td>
<td>48.3</td>
<td>58.9</td>
</tr>
<tr>
<td>Mycoplasma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hyopneumoniae</td>
<td>X06842</td>
<td>Gpl</td>
<td>2903</td>
<td>44.2</td>
<td>50.8</td>
<td>57.0</td>
</tr>
<tr>
<td>Mycoplasma pneumoniae</td>
<td>X06822</td>
<td>Gpl</td>
<td>2905</td>
<td>45.5</td>
<td>52.4</td>
<td>57.5</td>
</tr>
</tbody>
</table>

The lengths of the molecules from rhodobacters were corrected with respect to potentially instable insertions. Abbreviations: Dei, deinococci; Cya, cyanobacteria; Gph, Gram-positive bacteria with a high DNA G + C content; Gpl, Gram-positive bacteria with a low DNA G + C content; Pla, planctomycetes; Spi, spirochetes; Pal, Proteobacteria, α-subclass; Pbe, Proteobacteria, β-subclass; Pga, Proteobacteria, γ-subclass; Ttg, Thermotogales.
Fig. 2. Potential secondary structure model of the conserved core of 23S rRNA of Bacteria. The model is based on those published previously [41,43]. Uppercase letters in the consensus sequence indicate invariant bases, lower case letters indicate bases which are invariant in at least 90% of all available complete 23S rRNA sequences from Bacteria. All other positions are shown by open circles. Regions at which major insertions (more than 5 bases) occur are marked by arrows. The maximal number of inserted bases is given with the arrows (b, bases). Inserted helical structures are indicated by H followed by the corresponding number. Base numbering is that of the *Escherichia coli* 23S rRNA.
tions are essential for the recognition and alignment of homologous sequences, slowly evolving positions confer valuable information on the deepest groups, sites of higher variability are useful for the elucidation of more recent branchings. This has been shown for a region of bacterial 23S rRNA corresponding to positions 1521–1542 of the molecule of *E. coli* [36]. Conservation profiles (Fig. 1) have been established for four different phylogenetic levels i.e. the domain Bacteria, the Gram-positive bacteria with a low DNA G + C content, the phylogenetic group combining bacilli, lactobacilli, lactococci, streptococci, listeriae, staphylococci and others, and members of the genus *Staphylococcus*. The profiles have been calculated by defining the most frequent base at each individual position of an alignment of all available sequences from members of the particular group. They illustrate the different degrees of sequence conservation within that part of the sequence. The distribution of invariant and variable positions within the conserved core structure of bacterial 23S rRNA can be seen in Fig. 1. Fig. 2 is a potential secondary structure model (see below) of the conserved core of bacterial 23S rRNA based on the consensus sequence of all available complete 23S rRNA sequences from Bacteria.

**Insertions and deletions**

Within rRNAs there is apparently a bias towards base substitution [5]. Insertions and deletions, at least larger ones, are found especially in 23S rRNA genes and are restricted to predesigned regions [5]. Insertions and deletions are defined with respect to the 23S rRNA molecule of *E. coli*. Table 1 lists the lengths of 23S rRNAs from a selection of bacteria. The 5' and 3' termini have been defined in analogy to those of the *E. coli* molecule. The overall shortest molecules (2742 bases) are those of rhodobacters, whereas the molecule of *Mycobacterium leprae* contains the largest number of nucleotides (3132 bases). The regions at which remarkable insertions or deletions occur are marked in Fig. 2. There are stable as well as unstable insertions within 23S rRNA genes from Bacteria. Unstable insertions are excised during the maturation of the precursor rRNAs [37–39]. The processed 23S rRNA fragments are not religated, but the rRNAs remain fragmented within the ribosome. The lengths of the 23S rRNAs from the rhodobacters (Table 1) were corrected with respect to the putatively unstable insertions within their 23S rRNA genes [39]. The sum of the expected fragment length is given. A stable insertion of about 100 bases within domain III (helix 54a) [40,41] (Fig. 2) which is present within the mature ribosome has been described to be a common characteristic of the phylogenetic group of Gram-positive bacteria with a high DNA G + C content [42]. A major deletion of about 80 bases, again within domain III (Fig. 2), includes two secondary structure elements (helix 58 and 59) and seems to be a marker for representatives of the α-subclass of Proteobacteria. The deletion has not been found in any of the other phylogenetic groups. In general, there seems to be a tendency towards smaller 23S rRNA molecules among the α-subclass of Proteobacteria, whereas among Gram-positive bacteria with a high DNA G + C content large and small insertions are scattered over the molecule (unpublished results).

![Fig. 3. Potential secondary structure models of helix 21 of the 23S rRNAs from a selection of bacteria.](image-url)
Higher order structure

Ribosomal RNAs are capable of the formation of higher order structures (helices) by short- and long-distance intramolecular interactions of inverse complementary sequence stretches. Higher order structure models have been predicted based on the search for coordinated base changes at homologous positions in sequences from phylogenetically diverse origins, i.e. from only moderately related organisms [40,41,43]. The underlying assumption is that conserved function is reflected by structural similarity. Higher order structure elements are regarded as phylogenetically proven if the same or highly similar structures can be formed in rRNA molecules from different organisms, while the corresponding primary structure stretches differ. An example is given in Fig. 3. Potential secondary structure models of helix 21 of 23S rRNA (pos. 375–399 within the *E. coli* 23S rRNA) from a selection of bacteria are shown in comparison. Every base pairing is phylogenetically proven by (often coordinated) base changes. A limited but permanently increasing number of these intramolecular interactions has been experimentally proven for the *E. coli* molecule by crosslinking or nuclease protection studies [44].

23S rRNA molecules contain about 100 potential higher order structure elements, and there are about 50 within 16S rRNA molecules. Many of the helical and interhelical structure elements are highly conserved in length. A potential secondary structure model of the invariant core of bacterial 23S rRNA is shown in Fig. 2. The model is based on the consensus sequence redrawn from an alignment of all available complete 23S rRNA sequences from Bacteria. Positions which are invariant in all available 23S rRNA sequences from Bacteria are indicated by the corresponding symbols (A, C, G, and U), those which are invariant at least 90% of the sequences are shown as lower case letters, and all other positions are drawn as circles. Besides the conserved core regions, there is a great amount of variation within other parts.

---

**Fig. 4.** Alignment of homologous 23S rRNA sequence fragments according to primary structure similarity (A) and improvement of the alignment based on secondary structure similarity of helix 63 (B). Abbreviations: Bgl, *Bacillus globisporus*; Bli, *Bacillus licheniformis*; HNr, helix numbering [41,43]; a and b indicate the 5' and 3' halves, respectively; HSt, helix mask: positions which may be involved in secondary structure are indicated by ; Lde, *Lactobacillus delbrueckii*; Lla, *Lactococcus lactis*; Sca, *Staphylococcus carnosus*. Base paired nucleotides of helix 63 are indicated by bold italics in B. Asterisks mark (potentially) invariant residues.
of the structure. Sites of length variation are indicated by arrows. The maximal number of additional bases is given with the symbols.

**Phylogenetic analysis**

As outlined before, the different rates of change at different positions allow comprehensive phylogenetic analyses on the one hand, but complicate the analyses on the other hand, since the different positions cannot be regarded as equivalently informative if correct phylogenetic conclusions should be achieved.

**Sequence alignment**

The first step in molecular phylogenetic analyses is the proper alignment of the sequences. Phylogenetic trees derived from sequence data can only be as good as the underlying alignment. A correct alignment ensures that only homologous characters, i.e. residues which are derived from a common position within the ancestor molecule, are arranged in columns. They are compared for determining phylogenetic distances and reconstructing phylogenetic trees. Since invariant and conserved positions or regions are scattered along the sequence of rRNAs, the alignment of these regions is a straightforward procedure which properly can be done by computers. Within the variable and highly variable regions, especially if insertions or deletions are present, it is often difficult or even impossible to recognize homology from primary structure similarity. Furthermore, since similarity may be the consequence of common ancestry or of (multiple) changes, it does not necessarily indicate homology. Higher order structure analysis usually helps to resolve these problems. Within homologous higher order structure elements, 5' and 3' helix halves as well as internal or terminal single-stranded loops can be regarded as homologous elements and arranged in the alignment even if primary structures are different. In some cases, where higher order structure elements differ in size, it still remains difficult to optimally position alignment gaps. The determination of the quality of base pairing (for example, position and number of G–C pairs) provides additional information in some cases. Fig. 4 shows as an example homologous small 23S rRNA sequence fragments (E. coli 23S rRNA positions 1682–1768) aligned according to primary structure similarity (Fig. 4A) and the improvement of the alignment (Fig. 4B) considering higher order structure similarity (Helix 63; Fig. 5).

**Tree inferring methods**

The methods used for phylogenetic inference from sequence data have been reviewed recently [45–47]. The three major types of tree inferring approaches are pairwise distance [48], maximum parsimony [49], and maximum likelihood methods [50].

For the application of distance methods a matrix of pairwise dissimilarity values is calculated from the sequence alignment. Usually these dissimilarity values are transformed into phyloge-
nestic distances [46,51,52]. These corrections are used to compensate for superimposed substitutions. Since only contemporary sequences are available and there is only a limited number of possible character states, multiple base changes at a particular site cannot be recognized, and the real number of changes which have occurred until the evolution of the most recent organisms is definitely underestimated. The number of probably missed evolutionary events increases with the dissimilarity values. Based on distance matrices, phylogenetic trees can be reconstructed accord-

Fig. 6. Incorrect trees based on partial 23S rRNA sequences. (A) Distance matrix tree based on domain III sequence data. (B) Distance matrix tree based on domain IV sequence data. The bar indicates 10% expected base changes.
ing to described methods [45–52] as implemented in available programs [18,53]. These methods seek the tree for which, from topology and branch lengths, expected distances are most similar to those calculated from present-day sequences. A disadvantage of distance methods is that only overall dissimilarity values are used and all information about the behaviour of individual sequence positions is disregarded.

Maximum parsimony methods use the aligned sequences directly. The underlying model of evolution assumes that contemporary sequences were derived from their ancestors by acquiring a minimal number of changes [49]. The methods seek the most parsimonious trees among the possible tree topologies by determining the sum of changes which must have occurred to yield the sequences in the alignment. The tree(s) with the minimal overall length (the minimal number of changes needed) is (are) the most parsimonious tree(s). Parsimony methods tend to misplace organisms or groups if the amount of divergence is remarkably different in different lineages [54,55].

Maximum likelihood methods analyze the sequences on a site-by-site basis and incorporate an explicit model of sequence evolution [50]. This model may consist of the tree, an assumed mechanism of change (for example, changes are independent), and/or the probabilities of observing a change and its nature (for example, transition versus transversion) along each edge [45,46]. The parameters of the model can be derived from the actual data. That tree is the best tree which maximizes the congruence of model and data, i.e. the probability that the observed data would have occurred. The advantages of the maximum likelihood approach are that frequency and nature of changes at the individual sites are considered and that it can compensate for superimposed changes. However, because the method is computationally intensive, it has received little use up to now. Corresponding programs are implemented in the available PHYLIP package [53]. A new substantially faster version is available at the RDP [15].

All available treeing methods are based on assumptions (for example, that sites evolve independently) which may differ from the real event. Furthermore, the available programs usually do not perform exhaustive tests of all possible tree topologies. This is because of the enormous number of possible tree topologies and the expense in computing time. Thus, for 20 sequences there are approximately $8 \times 10^{20}$ possible trees. Most programs add the data according to their input order. Therefore, to see a potential influence of the input order on the tree, it is recommended to change it randomly and to perform serveral runs [45].

**Data selection**

As outlined above, rRNAs consist of independent domains, higher order structure elements, regions and positions which exhibit different degrees of evolutionary conservation. The rates of change at even homologous positions might be different in phylogenetically equivalent lineages or groups. Therefore, for phylogenetic analyses always complete or at least nearly complete sequences should be determined. In vitro amplification of rRNA genes and direct sequencing of amplified DNA provide easy access to almost complete rRNA sequences [41]. The analysis of partial sequences to reconstruct phylogenies may cause problems. Using only parts of the molecule, regionally clustered changes within molecules of a particular phylogenetic group may heavily influence the shape of the overall tree, and valuable information stored in other (not sequenced) parts of the molecule is not available.

Fig. 6 shows trees in comparison which had been derived from two sets of different arbitrarily chosen partial 23S rRNA sequence data from the same set of sequences from a selection of Gram-positive bacteria with a low DNA G+C content. Both trees have been reconstructed applying the same distance method with identical parameters. The tree in Fig. 6A is based on a 23S rRNA region comprising the rather variable domain III [41,43] of the molecule (Escherichia coli 23S rRNA positions 1275–1645), whereas the tree in Fig. 6B is based on the region comprising the more conserved domain IV (Escherichia coli 23S rRNA positions 1650–2010). Not only the branch lengths, which reflect the number of expected base changes, are different in both trees, but also the branching order is changed. The trees are
shown to demonstrate the potential effects of using inappropriate partial sequences and do not reflect phylogeny.

Obviously, complete sequences provide the maximum of available phylogenetic information. This does not mean that the inclusion of all positions gives optimal results. As discussed above, the diversity of rates of change at individual positions allows phylogenetic analyses over a wide range of phylogenetic distances. On the other hand, it complicates the reconstruction of reliable phylogenies. Invariant and conserved residues indicate homology and are useful for the alignment of the sequences, but confer little phylogenetic information. More variable positions define closer related groups, but are not informative for lower levels of phylogenetic relationships. Rapidly changing positions report on more recent events, but due to multiple base changes their inclusion in the analysis may add random noise especially concerning the deeper branches. Furthermore, within highly variable regions the alignment may not be optimal with respect to evolutionary history. Accurate phylogenetic inference requires the knowledge of variability, i.e. of the phylogenetic strength of the individual positions and their proper treatment. Maximum likelihood approaches use the information on positional variabilities by giving them different weights but do not remove them from the data set. The influence of these positions on the topologies of the trees can be recognized by performing several analyses while successively removing positions according to their degrees of conservation.

Subtrees can be rooted by including homologous sequences from only moderately related organisms as outgroup references. Different outgroup references should be tested to recognize artifacts which may be again due to ‘false’ sequence identities resulting from multiple base changes. Alternatively, consensus sequences derived from interrelated outgroup reference sequences may be used.

Besides the outgroup references, the addition of any new (even closely) related sequence to an existing data set influences the derived tree and may change its topology. In general, every new sequence which is different from those in the existing data set confers additional data on the information content of individual sequence positions as well as on the characteristics defining phylogenetic groups. The inclusion of this additional information may improve the tree locally or even globally. On the other hand, the addition of incomplete or incorrect sequences may produce artifacts and reduce the correctness of the tree. Lineages represented by a single organism can often not be stably positioned in phylogenetic trees.

A potential source of treeing artifacts applying distance and parsimony methods is compositional bias. The G + C contents of rRNAs in general do not vary within wide ranges (Table 1) but are usually higher in molecules from thermophilic bacteria [5]. However, the purine content is quite stable among bacterial rRNAs (Table 1). Therefore, transversion distance and transversion parsimony analyses should not be subject to these artifacts. Unfortunately, the information content of the sequences is reduced due to the two character state and, therefore, there is an increase of random identity. Maximum likelihood methods are in general less affected by the composition bias [5].

The significance of the relative branching order in a phylogenetic tree can also be tested by resampling methods [45,56,57]. In the jackknife method, a part of the data points (positions) is randomly dropped and a tree is reconstructed based on the randomly reduced data set. The procedure is repeated (at least 100 times, usually 1000 times) with alternatively truncated data sets. The significance of the individual branching points is expressed as the highest fraction of runs of recomputation in which the organisms or groups defined by the branching point appeared as monophyletic. An alternative method is bootstrapping, in which data points (positions) are randomly sampled for tree reconstruction. Some positions are sampled once, others several times, and some are not included at all.

**Conclusions**

Sequence analysis of rRNAs or of their genes has become a rapid standard technique and is no
longer the domain of specialists. It results in a rapidly growing database of phylogenetic as well as molecule-specific information. The correctness of the data (sequences) and their careful analysis are a prerequisite for the reconstruction of the most probable phylogeny of bacteria. However, biologists have to be aware that (even within large rRNAs) only a limited number of informative positions is available to elucidate four billion years of evolution. Only contemporary sequences are accessible, and each individual position carries the information on a short period of evolutionary time. Therefore, only a minor part of evolutionary events can be reconstructed. None of the methods for phylogenetic inference is perfect and none of the programs should be used as a black box and the results accepted without tests for significance. The application of the different methods outlined above on given data sets allows to test the robustness of inferred phylogenetic trees. In cases where lineages are branching in close proximity and the relative branching order does not remain stable when applying the different methods, this local topology cannot be regarded as significant. This should be indicated within the corresponding trees, e.g. as multifurcations.

**Bacterial phylogeny**

The detailed and comprehensive phylogenies inferred from 16S as well as 23S rRNA sequence comparisons provide three primary domains [14,58]: the Bacteria, Archaea and Eukarya. The domain of Bacteria comprises about 14 major groups (phyla) [5,14]. Fig. 7 shows for comparison both 23S rRNA and 16S rRNA based trees for the domain Bacteria. The trees have been reconstructed using all presently available complete (or almost complete) 23S rRNA sequences and almost equivalent data sets of 16S rRNA sequences. They were reconstructed using the same methods and applying the same parameters. The trees are drawn as consensus trees based on distance matrix trees corrected by including the results of parsimony and maximum likelihood analyses. Multifurcations indicate that branching orders could not be unequivocally determined. In both trees, thermophilic organisms (*Aquifex pyrophilus* and the *Thermotogales, Thermotoga maritima* and *Fervidobacterium islandicum*) keep the deepest branches. The green nonsulfur bacteria (*Chloroflexus aurantiacus* and *Herpetosiphon giganteus*) and the deinococci (*Thermus thermophilus* and ‘Deinonema’ spec.) are well separated in the 23S rRNA based tree and closer to one another in that based on 16S rRNA data. The remaining phyla are well defined in both trees, but a relative branching order could not be unambiguously defined using different treeing methods and parameters. In both trees the green sulfur bacteria (*Chlorobium vibrioforme* and *Chlorobium limicola*) are related to the cytophagas (*Flavobacterium odoratum, Flexibacter flexi-
The Gram-positive bacteria have originally been described as a monophyletic group [5,14] comprising two major clusters: the Gram-positive bacteria with a high genomic G + C content and the Gram-positive bacteria with a low genomic G + C content. However, the unity of the two groups is not significant in the 16S rRNA based tree. Based on the 23S rRNA data and comparative analysis of genes coding for the elongation factor Tu [20] the Gram-positive bacteria appear not to be of monophyletic origin. The branches within the 23S rRNA tree are generally longer than those of the 16S rRNA tree reflecting a higher degree of variability in 23S rRNA sequences. There is more phylogenetic information in 23S rRNAs, may be not in general, but at least for some phylogenetic levels.

In general 16S and 23S rRNA based trees are in good agreement. This also holds true for subtrees reflecting intraphylum relationships. The corresponding trees derived from rRNA sequence data from Gram-positive bacteria with a low DNA G + C are shown as examples in Fig. 8 [41].

**Bacterial identification**

As outlined above, rRNA sequences consist of differently conserved regions. There are elements which are invariant in sequence, size or structure in molecules of phylogenetically related organisms or groups but different in others. These structures can be regarded as signatures [14] for the respective organisms or groups and can be used for identification. There are for example single residues, base pairs, or combinations of one or both types defining the domains and lower groups at lower phylogenetic levels [14,59]. The presence of a characteristic insertion has been described as a signature for bacteria with a high DNA G + C content [42]. Partial sequences of appropriate parts of rRNAs containing signature elements can be used for identification of bacteria or their assignment to phylogenetic groups, but should not be used to reconstruct phylogenetic trees.

If the signatures are on the primary structure level, they may define diagnostic regions which provide potential target sites for specific hybridization probes [60]. Probes can be designed for groups of different phylogenetic depths. Probes have been designed and successfully applied for domains, subclasses, groups and species of bacteria [61–64]. However, the sequence variation of rRNA molecules is usually not sufficient to design strain-specific probes. Specific hybridization probes are in general short (15–20 bases) synthetic nucleic acids complementary to the target site. The probes are preferably designed complementary to the rRNA sequence since the higher amount (in comparison with genomic DNA) of rRNA molecules in living cells provide a natural target amplification. Applying
appropriate experimental conditions, stable probe-target hybrids are only formed if the hybrids are perfectly base paired. After the removal of non-bound probe, the hybrids can be detected by a probe-conferred label. A variety of commercial systems for labelling and detection are available. Various liquid and solid phase hybridization techniques are in use [60]. Organisms of the respective specificity groups can be identified in pure cultures and complex samples. For qualitative analyses the sensitivity of the techniques can be remarkably improved by in vitro amplification of target nucleic acids using the polymerase chain reaction technique (PCR) in combination with diagnostic or universal rRNA-specific oligonucleotide primers. In situ colony and cell hybridization techniques allow quantitative analyses [61,65,66]. At present, the most attractive technique is microscopic identification of bacteria after in situ cell hybridization with typically fluorescently labelled probes [61,66]. Besides the introduction of comparative rRNA sequencing for phylogeny inference, the development of in situ cell hybridization represents another landmark in the history of microbiology. For the first time, so far unculturable bacteria can be phylogenetically analyzed and identified in situ, even in complex natural samples [67–69]. The approach combines PCR-assisted rRNA sequence retrieval, phylogenetic analysis and specific detection of the corresponding bacterial cells with probes. The rRNA genes of a population are amplified in vitro and the different genes of the mixture subsequently singularized by cloning. After sequence determination they can be analyzed phylogenetically by comparison with the data base and used for the design of specific probes. In situ cell hybridization with these probes allows to associate the sequences with cellular morphotypes. These now phylogenetically known cells can be enumerated and their behaviour in the sample can be monitored.

Concluding remarks

After a short period of rapid evolution of methods for phylogeny inference and phylogenetically based identification, efforts have now to be strengthened to come to a worldwide coordination of the data and to define minimal standards for optimal data elaboration and interpretation. Databases of rRNA sequences aligned by specialists have already been established, but presently the frequencies of releases are far behind the nearly exponentially growing number of sequences. Methods for automation of sequence alignment, error check, signature analysis and phylogeny reconstruction are currently developed in our group, but still have to be improved. A database of probe sequences and additional information has also been initialized. Concerning the in situ cell hybridization technique, the sensitivity of the approach which so far depended on cells rich in ribosomes has to be improved. The power of the rRNA approach makes it reasonable that in the future all described bacterial species will be phylogenetically analyzed and that we will finally come to a genealogical classification. Furthermore, the new methods have now opened the door to investigation of bacterial diversity in various ecosystems and to identification and in situ detection of bacteria prior to their cultivation.

Acknowledgements

This work was supported by grants from the Commission of the European Union, the Bundesministerium für Forschung und Technologie, and the Deutsche Forschungsgemeinschaft.

References

12 Stackebrandt, E. (1991) Unifying phylogeny and pheno-
10 Fox, G.E., Pechman, K.J. and Woese, C.R. (1977) Com-
13 Kimura, M. (1983) The Neutral Theory of Molecular Evo-
17 Hori, It. and Osawa, S. (1979) Evolutionary change in
20 Ludwig, W., Weizenegger, M., Betzl, D., Leidel, E., Lenz,
19 De Rijk, P., Neefs, J.M., Van de Peer, Y., De Wachter, R.
18 Olsen, G.J., Overbeck, R., Larsen, N., Marsh, T.L., Mc-
16 Weisburg, W.G., Tully, J.G., Rose, D.L., Petzel, J.P., Oy-
9 Creti, R., Citarella, F. Tiboni, O., Samagelantoni, A.M.,
Palm, P. and Camarano, P. (1991) Nucleotide sequence of a DNA region comprising the gene for the elongation factor Tu from the ultrathermophilic archaea 
9 Creti, R., Citarella, F. Tiboni, O., Sanagelantoni, A.M.,
Palm, P. and Camarano, P. (1991) Nucleotide sequence of a DNA region comprising the gene for the elongation factor Tu from the ultrathermophilic archaea 
9 Creti, R., Citarella, F. Tiboni, O., Sanagelantoni, A.M.,
Palm, P. and Camarano, P. (1991) Nucleotide sequence of a DNA region comprising the gene for the elongation factor Tu from the ultrathermophilic archaea 


44 Ehrmann, M., Ludwig, W. and Schleifer, K.H. (1992) Specific species oligonucleotide probe for the identifica-


