MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, Dec. 1998, p. 1435–1491 Vol. 62, No. 4
1092-2172/98/$04.00+0

Protein Phylogenies and Signature Sequences: A Reappraisal of Evolutionary Relationships among Archaebacteria, Eubacteria, and Eukaryotes

RADHEY S. GUPTA*

Department of Biochemistry, McMaster University, Hamilton, Ontario L8N 3Z5, Canada

"The credible is, by definition, what is believed already,
and there is no adventure of the mind there."
Northrop Frye (74)

PREFACE

The recognition of archaebacteria as distinct life forms by Woese and coworkers in 1977 (256) has been hailed as one of the most significant developments in the history of microbiology and has profoundly influenced thoughts on the evolutionary relationships among living organisms. The discovery of this "third form of life" has led to the notion that prokaryotic cells are of two fundamentally different kinds, archaebacteria and eubacteria, and that of these, the archaebacteria are the closest relatives and direct ancestors of eukaryotic cells (Fig. 1a). The discovery of archaebacteria was initially based mainly on the 16S rRNA (oligonucleotide) sequences and phylogeny. However, during the past 10 years, much new information on different gene sequences, including the entire genomes of several prokaryotic and eukaryotic species, has accumulated (15, 26, 45, 66, 72, 73, 80, 119, 128, 138, 147, 215, 242). Based on these data, it is now possible to critically evaluate whether the three-domain proposal provides an accurate picture of the evolutionary relationship among living organisms or if a different type of relationship is warranted. The results of studies reviewed here indeed point to a very different evolutionary picture from the currently widely accepted one. In this review I present evidence based on molecular sequences that archaebacteria exhibit a

* Mailing address: Department of Biochemistry, McMaster University, Hamilton, Ontario, Canada L8N 3Z5. Phone: (905) 525-9140 ext. 22639. Fax: (905) 522-9033. E-mail: gupta@fhs.csu.McMaster.CA.

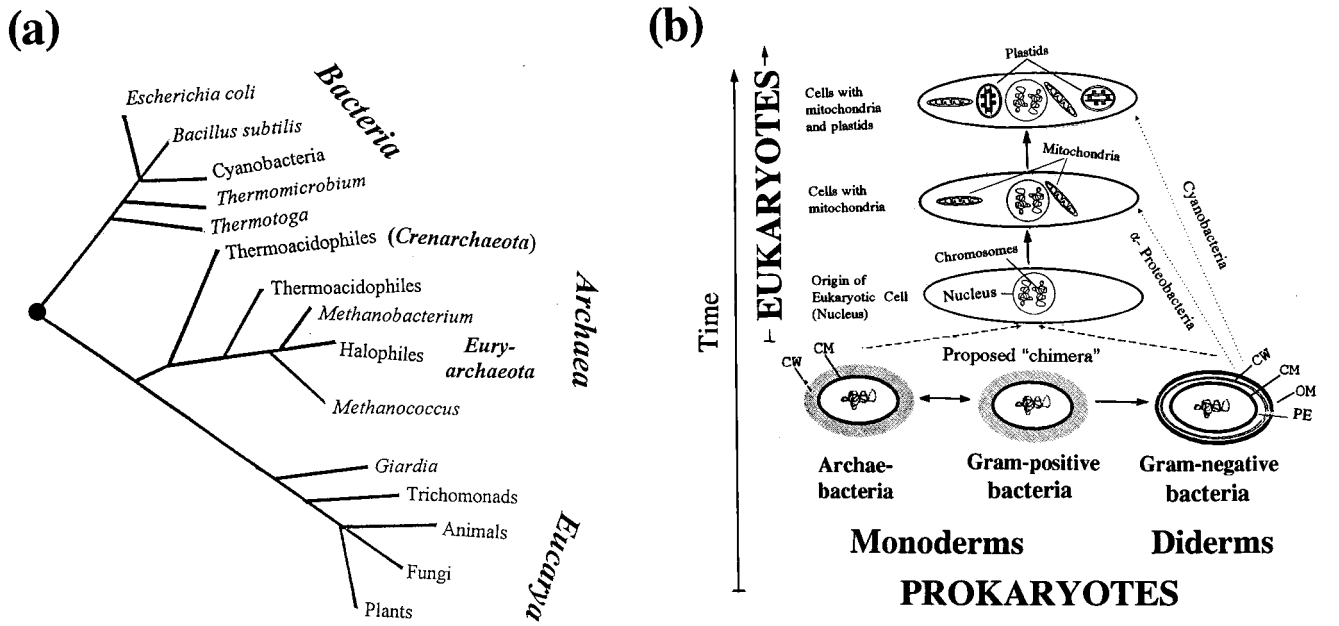


FIG. 1. Evolutionary relationships among living organisms in the three-domain model of Woese et al. (258) (a) and as suggested here based on protein sequence data and structural characteristics of organisms (b). In panel b, the solid arrows identify taxa that evolved from each other in the directions shown by accumulation of mutations and the dotted lines denote symbiotic events that led to the acquisition of mitochondria and plastids. These latter events, which are common in both models, are not shown in panel a. In panel b, the double-headed arrow between archaebacteria and gram-positive bacteria indicates the polyphyletic relationship between these groups for several genes. The terms "monoderm" and "diderm" refer to prokaryotic cells that are bounded by only one membrane or two different (cytoplasmic and outer) membranes, respectively. The dashed lines indicate the first fusion between an archaebacterium and a gram-negative bacterium that is postulated to have given rise to the ancestral eukaryotic cell (102, 105). Abbreviations: CM, cytoplasmic membrane; CW, cell wall; OM, outer membrane, PE, periplasm.

close and specific relationship to gram-positive bacteria and that the primary division within prokaryotes is not between archaebacteria and eubacteria but, rather, between organisms that have either a monoderm cell structure (i.e., prokaryotic cells surrounded by a single membrane, which includes all archaebacteria and gram-positive bacteria) or a diderm cell structure (i.e., prokaryotic cells surrounded by an inner cytoplasmic membrane and an outer membrane, which includes all true gram-negative bacteria) (Fig. 1b) (100). The sequence data also strongly indicate that the ancestral eukaryotic cell is not a direct descendant of the archaebacterial lineage but is a chimera that resulted from a unique fusion event involving two very different groups of prokaryotes—a thermoacidophillic archaebacterium (monoderm) and a gram-negative eubacterium (diderm), followed by integration of their genomes. Thus, all eukaryotic organisms, including the amitochondriate and aplastidic cells, received and retained gene contributions from both lineages.

CURRENT EVOLUTIONARY PERSPECTIVE

The quest for an understanding of the evolutionary relationships between extant organisms has posed a major challenge to biologists for centuries (23, 43, 159, 167). Since all living organisms are specifically related to each other by the presence of numerous common (or related) biomolecules and follow a similar complex strategy for growth and propagation, there is now little doubt that they all evolved from a common (universal) ancestor (3, 228). However, discerning how different major groups of organisms are related to each other and tracing their evolution from the common ancestor remains controversial and unresolved. After the invention of the microscope in the 17th century, studies on the morphological characteristics of cells from extant organisms led to the identification of two distinct types of cells (3), later termed prokaryotes and eukaryotes (34, 173), which could be readily distinguished. The eukaryotic cells are distinguished from prokaryotes by a number of different characteristics including the presence of a cytoskeleton, endomembrane system, etc. (3, 159). However, the hallmark feature of all eukaryotic cells is the presence of a membrane-bounded nucleus, and any organism lacking a nuclear membrane is considered a prokaryote (4, 34, 173). Eukaryotic organisms were classified into a number of different groups or kingdoms, namely, Animalia, Plantae, Fungi, and Protoctista, based on their detailed and complex morphologies and with the aid of fossil records (164, 248). However, a similar Linnaean approach to classification based on cell shape, physiology, and other characteristics was unsuccessful in detecting the phylogeny of prokaryotic organisms (23, 24, 121, 140, 194, 195, 227, 228, 230, 245, 250, 252, 254). The problem was partly due to their very simple morphologies but was also due in large part to the difficulty in determining which of the cellular features and characteristics of prokaryotes is most meaningful for taxonomic purposes.

Despite the ill-defined state of bacterial taxonomy, one empirical criterion that has proven of much practical value in the classification/identification of prokaryotes is their response to the Gram stain (121), discovered by Christian Gram in 1884 (88). As has been noted by Murray, "Gram-positiveness and Gram-negativeness are still unassailable characters except in Archaebacteria, the radiation-resistant cocci and . . . the wall-less mollicutes" (175). Gram staining involves successive treatment of cells with the basic dye crystal violet followed by treatment with iodine solution and then extraction with a polar organic solvent such as alcohol or acetone. The cells which resist decolorization and retain the blue-black dye complex are referred to as gram positive, whereas those which do not retain the stain are classified as Gram negative (12, 13, 88, 121). The

Gram-staining response, although not always reliable due to its dependence on cell physiology and cell integrity (11, 228), thus divides prokaryotes into two main groups, the gram-positive and the gram-negative (121, 228). Although the Gram reaction is an empirical criterion, its basis lies in the marked differences in the ultrastructure and chemical composition of the cell wall (14, 192, 228, 229, 235). The Gram-positive bacteria in general contain a thick cell wall (20 to 80 nm) that is very rich in cross-linked peptidoglycan (accounting for between 40 and 90% of the dry weight) and also containing teichoic acids, teichuronic acid, and polysaccharides (6, 14, 192, 229). Because of their rigid cell walls, these bacteria have been named *Firmicutes* in *Bergey's Manual of Systematic Bacteriology* (174); a number of other bacteria which possess the above structural characteristics but may show gram-variable (or gram-negative) staining are also placed in the same group. In contrast, all "true" gram-negative bacteria, named *Gracilicutes* in *Bergey's Manual* (174), have only a thin layer of peptidoglycan (2 to 3 nm) and have, in addition to the cytoplasmic membrane, an outer membrane containing lipopolysaccharides, which lies outside of the peptidoglycan layer. As noted by Trüper and Schleifer (244) "A clear separation of the Gram-positive and Gram-negative bacteria can be obtained by the differences in the ultrastructure and chemical composition of the cell wall". In the present work, I have used the term "gram negative bacteria" to describe prokaryotes whose envelopes contain a cytoplasmic membrane, a murine cell wall, and an outer membrane rather than by their Gram-staining response.

Based on the nature of the bounding layer of the cells, which is reflected in the Gram-staining reaction, a major microbiology textbook (228) suggested the division of prokaryotes into three main groups: "The Mycoplasma which do not synthesize a cell wall, the membrane serving as the outer bounding layer; the Gram-positive bacteria, which synthesize a monolayered cell wall; and the Gram-negative bacteria, which synthesize a cell wall composed of at least two structurally distinct layers." Although they could not know the extent of the problem, many earlier bacteriologists recognized the importance of cell structure and the bounding layer in the classification of prokaryotes: "It is self evident that the shape of the cell is of outstanding importance for determining the place of bacterium in any phylogenetic system" (140). However, as noted in a leading textbook, distinguishing between cells containing different types of envelopes was not an easy task (228): "The Gram-staining procedure is not always a wholly reliable method (and) the differentiation of these two subgroups (i.e., Gram-positive and Gram-negative) by other and more reliable methods is not easy; it requires either electron microscopic examination of wall structure in thin sections of the cells or chemical detection of the group specific polymers." In view of these difficulties, the results obtained were often difficult to integrate into a coherent scheme (24, 121, 174, 194, 195, 227, 228, 230, 245).

By the late 1950s and early 1960s, when microbiologists were feeling increasingly frustrated in their attempts to understand the natural relationship among prokaryotes, the era of molecular biology dawned. With this came the important realization, spelled out clearly by Zuckerkandl and Pauling (264), that the linear sequences of bases and amino acids in nucleic acids and proteins are informative documents containing a record of organismal evolutionary history from the very beginning and that in this regard the prokaryotic organisms are just as complex and informative as any eukaryote (65, 264). Thus, a comparison of sequences of the same gene or protein from various species could be used to deduce and reconstruct the evolutionary history of organisms. This marked the beginning of the field of molecular evolution. The rationale for using molecular sequence data to deduce the evolutionary relationship between organisms is described in a number of excellent reviews (58, 60, 61, 64, 65, 178, 236) and is not covered here except for certain relevant points.

The initial molecular approaches based on DNA base composition, nucleic acid hybridization, and immunological cross-reactivities were of limited use and were generally successful in establishing or rejecting relationships only among bacteria that were thought to be closely related species (224, 226, 228). The full impact of the molecular approach on evolutionary biology did not become evident until Woese and coworkers (71, 250, 256) had completed systematic studies of a significant number of living organisms based on the small-subunit rRNA sequences (SSU or 16S rRNA). The earlier studies in this regard were based on comparison of the oligonucleotide catalogs of the 16S rRNA, but these were later supplanted by phylogenetic analysis based on complete sequences of the molecules. These studies revealed that, based on genetic distances and signature sequences in the 16S rRNA, various prokaryotic and eukaryotic organisms fell into three distinct groups (71, 250, 256). One group consisted of all eukaryotic organisms, the second consisted of all commonly known bacteria (the term "eubacteria" was suggested for this group) including various genera of gram-positive and gram-negative bacteria and cyanobacteria, and the third group consisted of a number of previously little-studied prokaryotes (methanogens, extreme thermoacidophiles, and extreme halophiles) which grow in unusual habitats. Because of their assumed antiquity, this last group of prokaryotes was named "archaebacteria" (256).

In terms of their genetic distances (or similarity coefficients from oligonucleotide catalogs) based on rRNA, the archaebacteria were no more closely related to the eubacteria than to the eukaryotes. This observation, in conjunction with a number of unique characteristics of archaebacteria (e.g., lack of muramic acid in cell walls [127)] membrane lipids that contain ether-linked isoprenoid side chains [127, 133)], distinctive RNA polymerase subunits structures [263], and lack of ribothymine in the TΨC loop of tRNA), led Woese and collaborators to propose that the archaebacteria were totally distinct from other bacteria and constituted one of the three aboriginal lines of descent from the universal ancestor (71, 250, 256). The prokaryotes thus consisted of two distinct and non-overlapping (i.e., monophyletic) groups: eubacteria and archaebacteria, which were no more specifically related to each other than either was to the eukaryotes (250, 256). Since microbiology at the time was lacking any formal basis for phylogeny, this proposal, based on more defined and quantitative molecular characteristics, was generally favorably received, and within a decade most microbiology textbooks took notice of or were revised in the light of these new findings (6, 8, 14, 121, 192, 229).

The archaebacterial proposal received a major boost in 1989 when the phylogenies based on a number of protein sequences were added to the analysis, including those for the protein synthesis elongation factors EF-1α/Tu and EF-2/G, RNA polymerase subunits II and III, and F- and V-type ATPases (82, 126, 196). These studies again supported the distinctness of archaebacteria from eubacteria. Further, in contrast to the rRNA phylogeny, where only an unrooted tree was possible for archaebacteria, eubacteria, and eukaryotes, for the paralogous pairs of protein sequences (namely, EF-Tu and EF-G; and F- and V-ATPases) which appeared to be the results of ancient gene duplication events in the common ancestor of all extant life, it was possible to root the universal tree by using one set of genes as an outgroup for the other (82, 126). These studies indicated that the root of the universal tree lay between ar-

chaebacteria and eubacteria, and in both cases the eukaryotes were indicated as specific relatives of archaebacteria (82, 126). In 1990, Woese et al. (258) adopted this rooting, and a formal three-domain proposal for the classification of organisms was put forward. The proposal assigned each of the three groups, archaebacteria, eubacteria, and eukaryotes, a Domain status (a new highest taxonomic level) and renamed them *Archaea*, *Bacteria*, and *Eucarya*. The name *Archaea* was specifically proposed to indicate that this group of prokaryotes bear no specific relationship to the other prokaryotes (i.e., *Bacteria* or eubacteria) (258). This rooted version of the universal tree (Fig. 1a), commonly referred to as the archaebacterial or three-domain tree, is now widely accepted as the current paradigm in the field (54, 91, 171, 187, 258).

But does this tree or view represent the true relationship between the organisms? In recent years, much new information based on a large number of gene and protein sequences, including the complete genomes of several prokaryotic and eukaryotic organisms, has become available (26, 45, 66, 72, 73, 80, 119, 128, 138, 147, 215, 242). Based on this information, it is now possible to critically evaluate the three-domain proposal and its various predictions and to determine if this view is supported by all data or is true only for a subset of gene and protein sequences. These studies should also indicate whether a different sort of relationship between the organisms is more consistent with most of the available data. Since most biologists are not familiar with the assumptions and pitfalls of phylogenetic analyses, I will try to point out the strengths as well the subjective and weak aspects of such analyses so that the readers can understand and evaluate the results which form the bases for any classification.

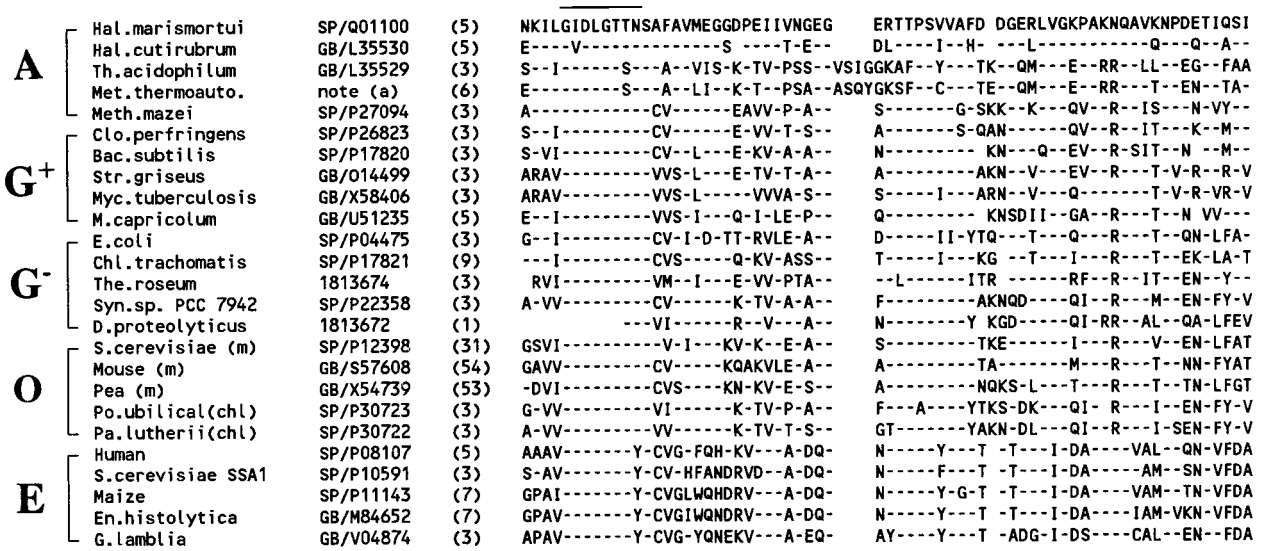
MOLECULAR PHYLOGENIES: ASSUMPTIONS, LIMITATIONS, AND PITFALLS

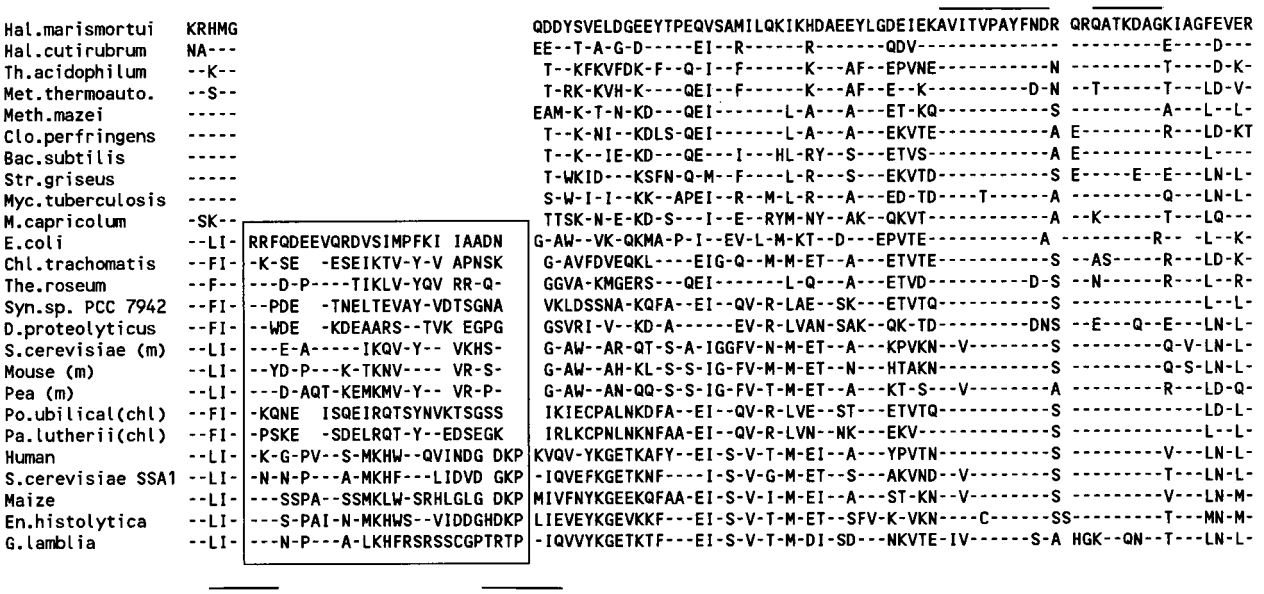
The use of molecular sequences for phylogenetic studies is based on the assumption that changes in gene sequences occur randomly and in a time-dependent manner and that a certain proportion of these become fixed in the molecules (58, 65, 136, 178, 236). The accumulation of changes in gene sequences in a quasi clock-like manner has given rise to the concept of "evolutionary clock" or molecular chronometer (136). Following the clock analogy (252), just as different hands or features (e.g., the month, day, minute, and second) in a clock move at very different rates, the changes in different gene sequences (or sometimes within different parts of the same gene) also occur at vastly different rates. Thus, some sequences which change very slowly (like the year, month, or day) are well suited for monitoring ancient events, while others, with a higher rate of change (like the hour, minute, or second), provide the sensitivity and resolution to measure relatively recent occurrences. Since the evolutionary history of life on this planet spans a vast period (approximately 3.8 Ga, 10^9 years), different sequences have different utilities in evolutionary studies. In the present context, where our main focus is on examining very ancient evolutionary events (e.g., relationships within the higher prokaryotic taxa and the origin of eukaryotic cells), the sequences which change very slowly and hence show a high degree of conservation in all extant organisms (i.e., the best-preserved molecular fossils) are most useful.

Phylogenetic analysis can be carried out based on either nucleic acid or protein sequences. For noncoding sequences such as various rRNAs, tRNAs, and introns, phylogenetic analysis can be carried out based on only the nucleotide sequence data. However, for gene sequences that encode proteins, analyses can be performed based on either the nucleic acid or the amino acid sequence data. For proteins, the two kinds of analyses appear analogous at first. In fact, the analysis based on nucleic acid sequences, with three times as many characters, would seem to be more informative (181, 250). While this is true in principle, for phylogenetic analyses involving distantly related taxa the increased information content in nucleic acid sequences as opposed to protein sequences is merely an illusion and in most cases is a major liability. The main reason for this lies in the degeneracy of the genetic code. All but two amino acids (Met and Trp) are encoded by at least two codons which differ in the third position. In view of this degeneracy, most changes in the third codon positions are selectively neutral (i.e., they do not result in any change in the protein sequence) and, as a consequence, change frequently even in closely related species (58, 60, 136). In distantly related taxa, which diverged from each other a long time ago, the bases at the third codon positions may have changed so many times that the actual bases found at these positions are random in nature and their information content is virtually nil. The inclusion of such bases in the analyses, therefore, would lead to uncertainty at every third position, thereby reducing the signal (i.e., positions which are evolutionary important)-to-noise (i.e., positions or changes which provide no evolutionary information) ratio in the data set.

Another important factor affecting the usefulness of nucleic acid sequences compared to protein sequences relates to the differences in the genomic G+C content of species (113, 231). The G+C content of different species is known to differ greatly (this is often true for two species within the same genus as well), and it is generally homogenized over the entire genome. In the protein-coding sequences, these differences in the G+C contents are accommodated by selective changes (i.e., codon preferences) in the third codon positions. The species which are rich in G+C show a strong preference for codons that have G or C in the third position (often >90%), whereas species with low G+C content predominantly utilize the codons with A or T in these positions. Thus, two unrelated species with similar G+C contents (e.g., either very high or very low) may have very similar bases in the third codon positions. If phylogenetic analysis is carried out based on nucleic acid sequences, these species may show a strong affinity for each other but for the wrong reason (113, 231). Thus, the third codon positions, rather than being informative, can introduce major bias into the analyses. For a similar reason but to a lesser extent, the bases in the first codon positions are also evolutionarily less informative and can cause reduced signal-to-noise ratio. Thus,

FIG. 2. Alignment of representative Hsp70 sequences from archaebacteria (A), gram-positive bacteria (G^+), gram-negative bacteria (G^-), eukaryotic-organellar (O), and eukaryotic nuclear-cytosolic (E) homologs. Small regions from the N- and C-terminal ends, which are not properly aligned in the global alignment of sequences and hence are not included in phylogenetic analyses, are not shown. The dashes indicate identity to the residue in the top line. The accession numbers of the sequences are shown. The boxed region shows the large insert in the N-terminal region present in all gram-negative bacteria and eukaryotic homologs. The solid lines above the sequence alignment identify several highly conserved regions that have proven useful to design degenerate primers for cloning purposes (57, 76, 102, 103, 107). The numbers at the beginning and at the end of the alignment denote the positions of the first and last amino acids included in individual protein sequences. The sequences were aligned by using the CLUSTAL program from PC Gene software package (IntelliGenetics), and minor changes were made to correct any visible misalignments. The abbreviations (m) and (chl) identify mitochondria and chloroplasts.





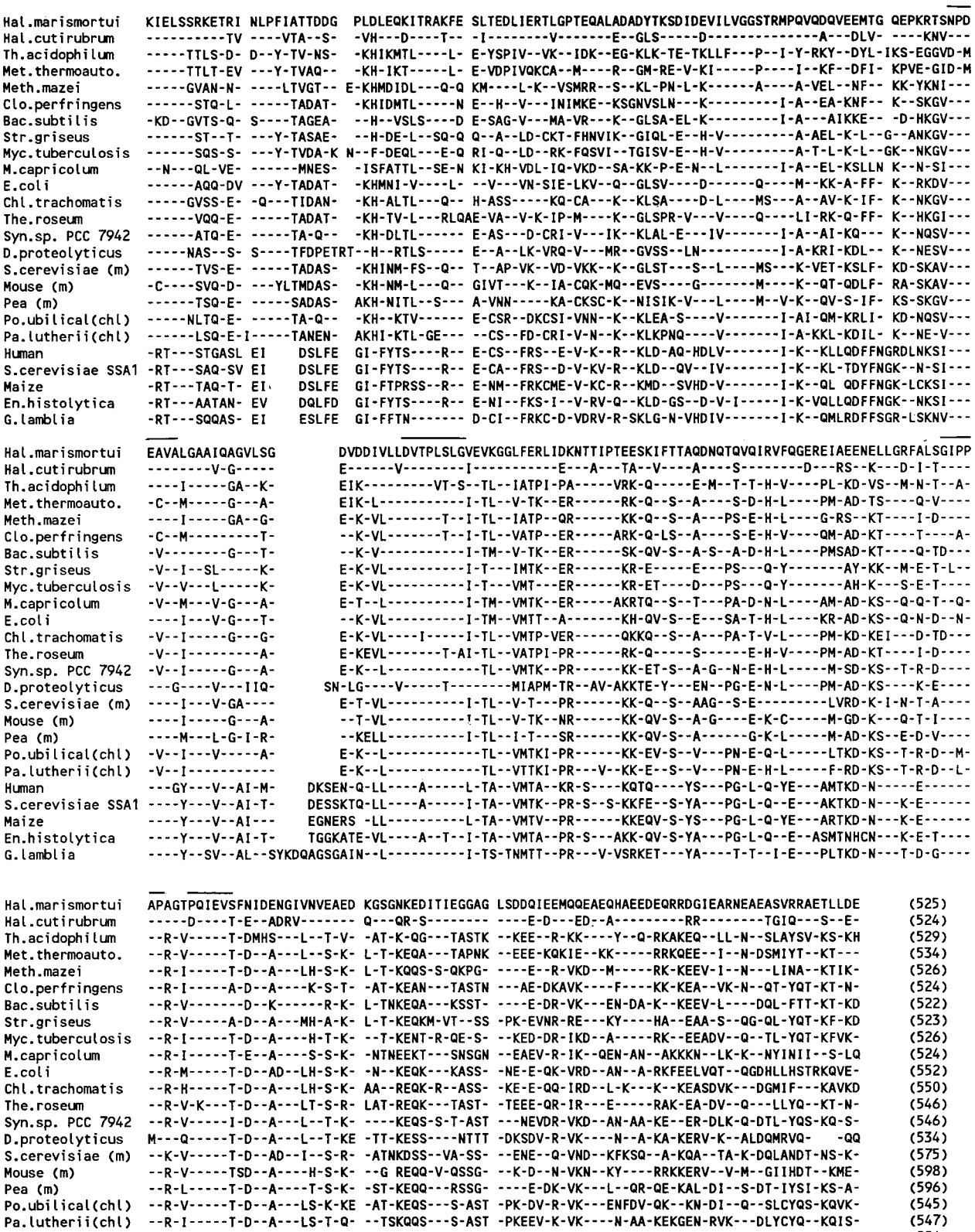


FIG. 2—*Continued.*

in the phylogenetic analyses of distantly related taxa with varying G+C contents, the larger number of characters in the nucleic acid sequences does not offer any real advantage, and if the bases at the third codon positions (and often those at the first positions as well) are not excluded from the analyses, misleading results could be obtained. In view of these considerations, for the protein-coding regions, the amino acid sequences, which are minimally affected by the differences in the G+C contents of the species, have proven more reliable and are the preferred choice for phylogenetic analyses (111, 113, 231).

In contrast to the protein-coding regions, where the codon degeneracy provides a natural mechanism for accommodating changes caused by G+C drifts, the effect of varying G+C compositions on structural nucleic acid sequences such as rRNA or tRNA remains largely undetermined. Thus, when comparing sequences from different species with varying G+C compositions, it is difficult to distinguish between the changes that are due to G+C drift (evolutionarily not significant) from those that are evolutionarily important. Thus, in any analyses based on structural nucleic acid sequences, the signal-to-noise ratio is inherently low. The effect that this will have on phylogenetic reconstruction cannot be easily determined or corrected, but this is a major and continuing source of concern in phylogenetic studies based on structural nucleic acids such as the 16S rRNA. As pointed out by Woese (251), "The problem (of) disparity in base composition is far more troublesome than is generally recognized and has almost received no attention to date. . . . It is important to understand the extent to which the general pattern reflects rRNA compositional disparity rather than the true phylogeny."

Another major problem in phylogenetic analyses is the reliability of the sequence alignment. The alignment of homologous positions in a set of sequences is the starting point in phylogenetic analyses from which all inferences are derived. Hence, the importance of having a reliable alignment for phylogenetic studies cannot be overemphasized. Most sequence alignment programs work by recognizing local similarity in different parts of molecules and then creating an alignment of all positions which maximizes the number of matches between the sequences, keeping the number of gaps introduced to a minimum (117). Although the alignment programs work similarly for both nucleic acid and protein sequences, there are important differences. In nucleic acid sequences there are only four characters, and hence the number of matches between any two sequences (unrelated) is expected to be a minimum of 25%; with the introduction of a small number of gaps, it is commonly in the range of 40 to 50%. In view of this, the probability of chance alignment of nonhomologous regions in two sequences is quite high, particularly if the sequences being compared are of different lengths and have either unusually high or low G+C contents. In contrast, in proteins each character has 20 states, which greatly reduces the probability of chance alignment between nonhomologous regions. There are no standard criteria for a good alignment, but it is generally assessed empirically by means of visual inspection. If the set of sequences contains highly conserved regions dispersed throughout the alignment, the proper alignment of such regions in all sequences is indicative of a good alignment. However, for sequences which do not contain many such regions, it is often difficult to get a reliable alignment for phylogenetic studies. Very often, differences in sequence alignment, the regions included in the phylogenetic analyses, or even the order in which the sequences are added in an alignment (151) could lead to important differences in the inferences drawn (42, 112).

Most extensive phylogenetic studies of living organisms have been carried out based on the SSU rRNA sequences (8, 77, 86, 149, 152, 224), which have been called the "ultimate molecular chronometers" by Woese (250). However, the alignment of rRNA sequences from various prokaryotic and eukaryotic species presents unique problems. In view of the large differences in the lengths of prokaryotic (≈1,500 nucleotides) and eukaryotic (≈2,000 nt) SSU rRNAs (mitochondrial SSU rRNA from some species is only 612 bp long [89]) and the wide variations in the G+C contents of species, a reliable alignment of rRNA sequences from distantly related taxa cannot easily be obtained based on the primary sequence data alone. The approach taken to get around this problem is to rely on the secondary-structure models of rRNA, based on the assumption that the secondary structure of the rRNA is highly conserved and provides a reliable guide for identification of homologous positions (252, 257, 259). Based on this, portions of the folded molecules (i.e., particular loops or stems) that are postulated to be similar in different sequences are aligned and used for phylogenetic studies.

The use of secondary-structure models for identification and alignment of homologous positions in the SSU rRNA is a very serious and far-reaching assumption. From an energetic point of view, the SSU rRNA can assume many different but equally likely secondary structures (259). While the proposed structures of rRNAs are supported by enzymatic digestion and chemical modification studies of some species (257, 259), their validity in distantly related prokaryotic and eukaryotic taxa is far from established. The effect that these far-reaching assumptions, on which all rRNA alignments are based (8, 33, 181, 184, 189, 224, 251), will have on the deduced phylogenetic relationships remains to be determined. However, it is clear that these assumptions have the potential to profoundly influence the outcome of any analyses (111).

In contrast to the rRNA sequence alignment, alignment of amino acid sequences of a highly conserved protein such as the 70-kDa heat shock chaperone protein (Hsp70) requires minimal or no assumptions. Because of the similar size of this protein in various prokaryotic and eukaryotic species (including organellar homologs) and its high degree of sequence conservation, a good alignment of the sequences from various species is readily obtained by using any common sequence alignment program (117) or even manually by placing the sequences next to each other. Figure 2 shows an alignment of 25 Hsp70 sequences covering the prokaryotic and eukaryotic spectrum as well as organellar homologs. The alignment shown was obtained with the CLUSTAL program from the PCGENE software, and only minor corrections to it have been made manually. The large number of identical and conserved residues present throughout the length of this alignment gives confidence that the observed alignment is reliable. The global alignment of Hsp70 sequences shows many regions that are nearly completely conserved in all species. Degenerate primers based on these sequences have been successfully used to clone the gene encoding Hsp70 from a wide range of prokaryotic and eukaryotic organisms (56, 57, 76, 102, 103, 107, 108).

Once a (reliable!) sequence alignment has been obtained, three main types of methods are used for phylogenetic reconstruction: those based on maximum parsimony (58, 64), those based on pairwise genetic distances between the species (65, 207), and the maximum-likelihood method (58, 137). These methods interpret the sequence alignment in different ways, and therefore the results obtained from them often differ (110, 238). All these methods, as well as the others (e.g., evolutionary parsimony [152]), can give rise to incorrect relationships under different conditions. Five main factors affecting the out-

come of these analyses are (i) an underestimation of the number of genetic changes between the species (often multiple changes in a position are counted as either one or no change); (ii) the long-branch-length effect, where two distantly related taxa may appear more closely related than they truly are if there are no intermediate taxa to break the long branches (62); (iii) large differences in the evolutionary rates among different species in the data set; (iv) horizontal or lateral gene transfers between the species (236a); and (v) comparison of paralogous sequences which are the results of unidentified ancient gene duplication events (62, 110, 152, 233, 238). In most cases, it is difficult to ascertain the effects of different factors and to determine which phylogenetic method is more suitable or reliable. Hence, phylogenetic analyses are generally carried out by different methods to see if all the methods give similar results.

The reliability of phylogenetic relationships inferred from the above methods is commonly assessed by performing a bootstrap test (59). In this test, the aligned sequences are sampled randomly and certain numbers of columns in the original alignment are replaced with columns from elsewhere in the sequences to obtain 100 or more different alignments, each containing the same number of columns. Thus, in a given bootstrap set, some columns will not be included at all, others will be included once, and still others will be repeated two or more times. Phylogenetic analysis is then performed on each of the bootstrap replicates, and a consensus tree from this data is drawn. The main purpose that bootstrap analyses serve is to provide a measure of the variability of the phylogenetic estimate or confidence levels in the observed evolutionary relationships. If the sample data throughout the sequence length support a particular relationship, this will be reflected in the grouping of the species in all (or a vast majority) of the bootstraps. The results of these analyses are presented by placing bootstrap scores (indicated by the percentages or the number of times that different species group together in bootstrap trees) on different nodes in the tree. Bootstrap values of >80 to 85% are generally considered to provide good support for a specific phylogenetic relationship.

Despite due care in the alignment and analyses of the sequence data, interpretation of the phylogenetic trees that are obtained is not straightforward. The most common problem in this regard is that phylogenetic trees based on different genes or proteins may differ from each other in terms of the evolutionary information that they provide. Based on the clock analogy discussed above, some genes are better suited to resolve certain relationships than are others. Thus, while a particular relationship may be clearly resolved and strongly supported by one gene phylogeny, the same relationship may not be obvious from a different gene phylogeny. Such results are generally regarded as controversial by many scientists, including evolutionary biologists (49, 53, 69, 70), but it is important to realize that they are not. Part of the problem in the interpretation of new data stems from the commonly held perception that phylogenetic trees based on just one or two molecules (e.g., 16S rRNA) can clearly establish the evolutionary relationships between all extant species (181, 184, 188, 202, 224, 250–252). This means that any results that do not concur with the 16S rRNA phylogenies are generally considered deviant and suspicious (69). However, such a notion is clearly erroneous, in view of the limitations of the rRNA-based phylogenies noted above and the inability of the 16S rRNA trees to resolve the branching orders of the deeply lying taxa within eubacteria: "(In the 16S rRNA phylogeny) the majority of the bacterial phyla arise in such a tight radiation that their exact order of branching has yet to be resolved" (252).

Cognizant of these problems, many scientists working in this area have urged caution in the interpretation of phylogenetic data. Woese wrote (252): "The scientifically proper stance for the microbiologists to take at this juncture will be to treat these phylogenies (bacterial) as hypotheses, and test them using other molecules, phenotypic characteristics of the organisms, and so on. When the same or very similar relationships are given by different molecular systems or when new phenotypic similarities consistent with the projected phylogenies turn up, then that phylogeny can be confidently accepted"; Rothschild et al. wrote (202): "We encourage phylogenetic analyses where molecular approaches are evaluated in the light of other available data, and where the strengths as well as subjective and weak aspects of the analyses are made explicit"; and Murray et al. stated (177): "The integrated use of phylogenetic and phenotypic characteristics, called polyphasic taxonomy (38), is necessary for the delineation of taxa at all levels from Kingdom to genus". I do not think any evolutionary scientist will disagree with the above statements or suggested approaches.

It is clear from the above discussion that the results of phylogenetic analyses should not be uncritically accepted but instead should be evaluated in the light of other available data, including data from morphological, geological, and fossil sources. There is also a pressing need to develop additional sequence-based criteria for determining the evolutionary relationships among species, which are based on minimal assumptions and which could be readily understood and interpreted by both specialists and nonspecialists. In the next few sections, I present evidence that conserved inserts or deletions restricted to specific taxa (170), which are referred to as signature sequences in the present work, provide such criteria.

SEQUENCE SIGNATURES AND THEIR IMPORTANCE IN EVOLUTIONARY STUDIES

Signature sequences in proteins could be defined as regions in the alignments where a specific change is observed in the primary structure of a protein in all members of one or more taxa but not in the other taxa (99, 107, 198). The changes in the sequence could be either the presence of particular amino acid substitutions or specific deletions or insertions (i.e., indels). In all cases, the signatures must be flanked by regions that are conserved in all the sequences under consideration. These conserved regions serve as anchors to ensure that the observed signature is not an artifact resulting from improper alignment or from sequencing errors. Although changes of various kinds can serve as sequence signatures (56, 99), in the analyses presented here I have mainly considered only signatures involving indels. My reason for focusing on indels is that I think they are less likely to result from independent mutational events occurring over a long period (see below), compared with change in nucleotides and hence amino acids. Since this review is the first detailed attempt to use conserved indels as phylogenetic markers to discern the course of evolutionary history, a discussion of the rationale for such studies as well as their limitations and pitfalls is provided.

The rationale of using conserved indels in evolutionary studies could briefly be described as follows. When a conserved indel of defined length and sequence, and flanked by conserved regions (which ensure that the observed changes are not due to improper alignment or sequencing errors), is found at precisely the same position in homologs from different species, the simplest and most parsimonious explanation for this observation is that the indel was introduced only once during the course of evolution and then passed on to all descendants. This is a minimal assumption implicit in most evolutionary analyses.

Thus, based on the presence or absence of a signature sequence, the species containing or lacking the signature can be divided into two distinct groups, which bear a specific evolutionary relationship to each other. A well-defined indel in a gene or protein also provides a very useful milestone for evolutionary events, since all species emerging from the ancestral cell in which the indel was first introduced are expected to contain the indel whereas all species that existed before this event or which did not evolve from this ancestor will lack the indel. Further, if specific indels could be identified in proteins that coincide with or were introduced at critical branch points during the course of evolution, such signatures could serve as important phylogenetic markers for distinguishing among major groups of organisms.

In using conserved indels as phylogenetic markers, two potentially serious problems that could affect the interpretation of any data should be kept in mind. First, there is the possibility that the observed indel was introduced on multiple occasions in different species due to similar functional constraints and selection pressure rather than being derived from a common ancestor. Second, lateral gene transfer between species could also readily account for the presence of shared sequence features in particular groups of organisms. While a definitive resolution of the question whether a given sequence signature is due to common ancestry or results from these two causes is difficult in most cases, important insights concerning the significance of such data are often provided by consideration of information from other sources.

The most important and relevant information bearing on this issue is provided by consideration of cell structure and physiology. In this context, it should be emphasized that the aim of phylogenetic analysis is to explain and reconstruct the evolutionary history of organisms. Hence, the structural and physiological characteristics of organisms are of central importance, and they should be the ultimate arbiter in determining the significance of such data. Without this context, phylogenetic analysis of sequence data could become an end in itself, bearing little relation to the organisms. Therefore, if the inference derived from a given signature sequence or phylogenetic analysis is consistent with an important structural (e.g., cell envelope structure) or physiological attribute of the organisms, it is likely that we are on the right track, and it gives confidence in the correctness of the inference. On the other hand, if the inferences based on signature sequences and phylogenetic analyses are at a variance with important structural and physiological characteristics, one should ask questions about why it is so rather than distrusting or ignoring these characteristics.

Another useful criterion in assessing whether a given signature is of evolutionary significance is provided by its species distribution. If a given sequence signature is present in all known members of a given taxa, it is more probable that it was introduced only once in a common ancestor of the group and then passed on to all descendants. In such cases, phylogenies based on other gene sequences are also expected to be generally consistent with and support the inference drawn from the signature. In contrast, when a shared indel is present either in only certain members of particular taxa or when species containing the signature show no obvious structural or physiological relationship, the possibility that the observed signature is a result of independent evolutionary events or horizontal gene transfers becomes more likely. In our analysis, we have come across several examples of signature sequences which provide evidence of lateral gene transfers between species (unpublished results). Such signatures are of limited use in deducing phylogenetic analysis and, except for a few, will not be described here.

The presence of well-defined signature sequences in proteins should allow one to establish evolutionary relationships among species by means of molecular cladistic analysis. This approach, although not generally applicable to all proteins (because most proteins do not contain useful sequence signatures), has certain advantages over traditional phylogenetic analyses based on the gene or protein sequences. First, in traditional phylogenetic analysis, the evolutionary relationships among different species are determined based upon the assumption of a constancy of evolutionary rate in all species (58, 60, 65, 136). Since this assumption is rarely correct over long periods (84), the differences in evolutionary rates could lead to incorrect species relationships. However, the signature sequences, such as conserved indels of defined sizes, should not be greatly affected by the differences in evolutionary rates. The proteins which are greatly affected by the differences in evolutionary rates are unlikely to contain well-defined indels in conserved regions and hence will be excluded from consideration. A second common and serious source of problems in phylogenetic analysis involves sequencing errors, and anyone involved in DNA sequencing should be familiar with this. For example, sequence compressions which are not satisfactorily resolved are a common occurrence, particularly in G+C-rich sequences. The errors introduced in reading such regions could lead to either localized (from base and amino acid substitutions) or extended (from frameshifts) changes in the gene or protein sequences. In one study, the error frequency in DNA sequences in the databases has been estimated at 3.55% (146), although other estimates indicate it to be much lower (145). An additional but related problem involves the increasing number of sequences in the databases which have been obtained by PCR amplification and sequenced by automated means. The higher rates of sequence errors and contamination in such sequences should be a cause of concern. These factors could affect the branching orders of species in phylogenetic trees. However, it is highly unlikely that a sequencing error could give rise to an indel of a defined length and sequence at a precise position within a conserved region. A signature of even one amino acid involves the addition or deletion of three nucleotides in the DNA sequence at a precise position and hence is highly significant. Third, a very common problem in evolutionary analyses (discussed in the previous section) is that the phylogenetic trees based on certain genes (or proteins) may fail to resolve the branching orders (e.g., low bootstrap scores for the nodes) for particular groups of species and hence the results of these studies will be indeterminate; i.e., they neither support nor refute a particular relationship (21, 85). However, this is not a problem in the case of signature sequences, where the relationship is assessed based on the presence or absence of a given signature and thus its interpretation is unambiguous. One expects that the relationship indicated by signature sequences should generally be consistent with and supported by the phylogenetic analysis based on other gene or protein sequences. However, the analyses based on signature sequences are limited in one sense: whereas a phylogenetic tree provides information about evolutionary interrelationships among all species in a tree, a given signature sequence is limited to distinguishing and establishing the evolutionary relationship between the two groups of species, i.e., those containing and those lacking the signature.

ROOT OF THE PROKARYOTIC TREE: ANCESTRAL NATURE OF ARCHAEBACTERIA AND GRAM-POSITIVE BACTERIA

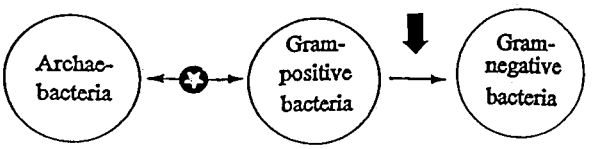
To fully understand and correctly interpret the implications of a given sequence signature, a reference point is required. When an indel is present in one group of species and absent from others, it is difficult to say a priori which of these groups is ancestral and which is derived. While this problem cannot be resolved in most cases, one instance where valuable additional information helpful in resolving this question is available corresponds to a signature identified in the Hsp70 family of proteins. Hsp70 homologs from different gram-negative bacteria contain a conserved insert of 21 to 23 amino acids which is not present in any homolog from gram-positive bacteria or archaebacteria (Fig. 3) (103, 107, 108). This sequence signature could result either from a deletion in the common ancestor of all archaebacteria and gram-positive bacteria or from an insertion in the common ancestor of all gram-negative bacteria. Depending upon which of these scenarios is correct, one of these groups of prokaryotes becomes ancestral and the other becomes derived. Resolution of this question is provided by a number of different observations.

First, based on the duplicated gene sequences for EF-1α/Tu and EF-2/G proteins, where one set of sequences could be used to root the other tree, the roots of both EF-1α/Tu and EF-2/G trees have been shown to lie between the archaebacterial lineage and the eubacterial species *Thermotoga maritima* (7, 21, 112). A tree for EF-1α/Tu sequences, which was rooted by using EF-2/G, is shown in Fig. 4. As seen in this figure, the root of the tree lies in between archaebacteria and eubacteria and the deepest branches within eubacteria consist of *T. maritima* and other gram-positive bacteria. A similar rooting of the universal tree in between archaebacteria and *T. maritima* has been independently made based on trees constructed from homologous isoleucine-, leucine-, and valine-tRNA synthetase sequences (20). Although the species *T. maritima* has been assumed to be a gram-negative bacterium in the past (184, 250, 251, 258), recent studies based on several proteins provide evidence that it should in fact be grouped with gram-positive bacteria (22). This inference is supported by signature sequences in Hsp70 (Fig. 3) and a number of other proteins (see "Evolutionary relationships among prokaryotes"), where *T. maritima* behaves similarly to various gram-positive bacteria and differently from different gram-negative bacteria. Phylogenetic analyses based on a number of proteins, i.e., Rec A (55, 131, 247) and sigma factor 70 (39), also provide evidence of a grouping of *T. maritima* with gram-positive bacteria. Most importantly, Cavalier-Smith (31) has pointed out that *T. maritima*, similar to other gram-positive bacteria, is bounded by only a single unit lipid membrane, which I consider to be the main defining characteristic of gram-positive bacteria. In view of these observations, the results of the above rootings indicate that the root of the prokaryotes lies between archaebacteria and gram-positive bacteria.

A second independent line of evidence supporting the ancestral nature of the clade consisting of archaebacteria and gram-positive bacteria is provided by a comparison of sequences for the Hsp70 and the MreB families of proteins. We have previously shown that MreB protein, which is about half the length of Hsp70 (about 340 amino acids [aa], with respect to 600 to 650 a.a. for Hsp70) and is present in all major groups of prokaryotes (archaebacteria, gram-positive bacteria, and gram-negative bacteria), shows significant similarity to the N-terminal half of Hsp70 sequences (107), where the large indel in the Hsp70 homologs is present. The three-dimensional structures of the MreB protein and the N-terminal half of Hsp70 are also very similar (18, 65a), supporting the view that these proteins have evolved from a common ancestor (18). Since both Hsp70 and MreB proteins are found in all main groups of prokaryotes, they very probably evolved by an ancient gene duplication in the universal ancestor, before Hsp70 acquired the C-terminal domain (104, 107). In view of this, we expect that if the above indel in Hsp70 is an insert in gram-negative bacteria, the MreB protein sequences should not possess it. On the other hand, if the homologs containing the insert are ancestral, this insert should also be found in the MreB sequences. A comparison of MreB and Hsp70 sequences from the major group of prokaryotes (Fig. 5) shows that, similar to the Hsp70 from archaebacteria and gram-positive bacteria, this insert is not present in any of the MreB sequences, including those from gram-negative bacteria. (It should be mentioned that since MreB and Hsp70 are very distant homologs, the sequence similarity between these proteins is limited. However, despite this fact, the inference that MreB protein does not contain the insert is quite apparent.) This observation provides strong independent evidence that the prokaryotic organisms lacking the insert (i.e., archaebacteria and gram-positive bacteria) are ancestral and that this insert was introduced into Hsp70 in a common ancestor of the gram-negative bacteria (104, 107). I will refer to this insert, which is a distinguishing feature of gram-negative bacteria and eukaryotes, as the diderm insert, signifying its point of evolutionary origin.

Lastly, the view that archaebacteria and gram-positive bacteria are ancestral lineages is also consistent with the available evidence concerning the planet's early environment. Based on Earth's geological history, the conditions under which the earliest organisms evolved were hot and anaerobic (Fig. 6). The widespread prevalence of the ability to exist under these conditions in various archaebacteria and gram-positive bacteria (67, 186, 232, 250) is consistent with the view that these groups are ancestral. Based on the above pieces of evidence, all of which lead to a similar inference, I am going to assume that the rooting of the prokaryotic tree between (or within) archaebacteria and gram-positive bacteria is correct, and I will examine whether this rooting can explain other observations and phylogenies.

The root provides an important reference point for evolutionary studies. By using this reference point, it should now be possible to understand and interpret signature sequences in different proteins to piece together the evolutionary relationship and history of the other groups of prokaryotes. In the following sections, I describe signature sequences in different groups of species and my interpretation of them based on the above rooting. Since a great deal of work that follows is based on signature sequences that are reported for the first time, it is appropriate to describe the approach taken to identify the signature sequences. The signature sequences in a number of proteins such as Hsp70 and Hsp60 were empirically discovered (96, 104, 107, 108). However, the complete genomes of several gram-positive bacteria (*Mycoplasma genitalium* [73], *Mycoplasma pneumoniae* [119], and *Bacillus subtilis* [147], gram-negative bacteria (*Haemophilus influenzae* [66], *Escherichia coli* [15], *Synechococcus* sp. strain PCC 6803 [128], *Helicobacter pylori* [242], *Borrelia burgdorferi* [72], and *Aquifex aeolicus* [45]), and archaebacteria (*Methanococcus jannaschii* [26], *Methanobacterium thermoautotrophicum* [215], and *Archaeoglobus fulgidus* [138]) have recently been reported. In view of this, to search for signature sequences in different proteins, a systematic approach was used. For these purposes, we performed a BLAST search (5) on each of the unique proteins identified in



				65		102
Monoderm Prokaryotes	A	Th. acidophilum	GB/L35529	NPEGTIFAAKRKMG	TDYKFKVFDKE	FTPQQISAFILQKI
		Hal. marismortui	SP/Q01100	--DE--QSI--H--	QD--SVELDGE-	Y--E-V--M-----
		Hal. cutirubrum	GB/L35530	--DQ--ASINAH--	EE--TVALGGD-	Y--EE---R-----
		Meth. mazei	SP/P27094	--DN-VYSI--H--	EAM--VTLNG-D	Y---E---M----L
		Met. thermoauto.	2622398	---N--T-I--S--	--R-V--HG--	Y---E---------
	G^+	Strep. pyogenes	1750267	-- E-VISI-S---	-SE-VSANG--	Y---E---M---YL
		Strep. agalactiae	1750261	-- D-VISI-S---	-SE-VSANG--	Y---E---M---YL
		Strep. pneumoniae	1750264	-- D-VISI-S---	-SE-VSANG--	Y---E---M---YL
		Strep. mutans	2145133	-- E--LSI-S---	-SE-VSANA--	Y---E---M---YL
		Myc. leprae	SP/P19993	-VDR--RSV--H--	S-WSIEIDG-K	Y-A-E---RV-M-L
		Myc. paratuber.	GB/X58406	-VDR--RSV--H--	--WSIEIDG-K	Y-A-E---RV-M-L
		Myc. tuberculosis	GB/Q00488	-VDR-VRSV--H--	S-WSIEIDG-K	Y-APE---R--M-L
		Str. coelicolor	GB/O14499	-VDR--RSV--H--	--W-VNLDG-D	-N-------V---L
		Str. griseus	SP/P17820	-VDR--RSV--H--	--W-IDLDG-S	-N---M-------L
		Bac. subtilis	GB/U39711	-- N--MSI--H--	----VEIEG-D	Y---EV--I---HL
		Bac. megaterium	GB/D30690	-- N--ISV--H--	--H-VEAEG-Q	Y---EM--I---HL
		Bac. stearotherm.	1568474	-- N-VISI--H--	----VEIEG-K	Y---E---I---YL
		Ery. rhusiopathiae	GB/L08201	-- NSAVSV--LI-	-GE-VTLEG-D	Y--EE---M--GYM
		Meg. elsdenii	Unpubl.	---N--SSI--H--	ES-TVDIQG-K	Y---E---M----L
		M. pneumoniae	1674090	-- N--VSI--L--	-SN-VT-KNPD	GSTKEL--Q--SYL
		M. capricolum	GB/U51235	-- NVVQSI-S---	-TS-VNLEG-D	YS-E----E--RYM
		M. genitalium	1361531	-- N--VSI--L--	-SN-V--TT--	LS-E-V--Q--SYL
		Clo. acetobutyl.	SP/P30721	--DK--ISI-----	-AE-VAID--N	Y---E---M----L
		Clo. perfringens	SP/P26823	--DK--MSI--H--	----VNIDG-D	LS--E---M----L
		L. lactis	GB/M98865	-- E--ISI-S---	-SE-VSANG--	Y---E---M---NL
		Sta. aureus	GB/X75428	-- N-VQSI--H--	----VDIEG-S	Y---E---M---NL
		T. maritima	Unpubl.	---R--KSI-----	----VRID---	Y---E------K-L
Diderm Prokaryotes	G^-	E. coli	SP/P04475	--QN-L--I--LI-	RRFQDEEVQRDVSIMPFKIIAAD	NG-AWVE-KGQK	MA-P----EV-L-M
		H. actinomycet.	1169375	--KN-L--I--LI-	---E----K--ID----A-TK--	NG-AWVE-KGEK	LA-P----EV-K-M
		H. influenzae	1542849	--KN-L--I--LI-	---ES------IK----E-TR--	NG-AWVN-KGDK	LA-P----EV-K-M
		F. tularensis	1352284	--DN-F--I--LI-	-KYD-KA--E-IKKV-YAV-K--	NG-AWVATKG-K	MA-P-V--EV-R-M
		Sal. typhimurium	1389758	--QN-L--I--LI-	----------Y---G--	NG-AWLD-KGQK	MA-P----EV-K-M
		Nit. europaea	2094746	---N-L--I--LI-	---DE-V--K-I-VT-Y--VR--	NN-AWIEARGRK	IA-PEV--QV-I-M
		Rhodopseudomonas sp.	2058265	---R-F--V--LI-	--YD-PM-EK-KGLV-Y--VK-S	NG-AWVEADG-T	YS-S-V-------M
		Rho. capsulatus	1373328	--TN-V--V--LI-	--TT-A--EK-KKLV-YN-VDGG	NG-AWVE-RGEK	-S-A-V--V----M
		Brad. japonicum	1769954	---R-F--V--LI-	--YD-PM-EK-KKLV-Y--VK-S	NG-AWVEADGQT	YS-S-V-------M
		Pse. cepacia	GB/L36603	--KN-L--V--LI-	---EEK---K-IGL--YS--K--	NG-AWVEGHGEK	MA-P-V--EAVR-M
		R. meliloti	GB/L36602	---N-L--I--LI-	-T-E-PTT-K-KGMV-Y--VK--	NG-AWVEAHGTS	YS-S----M----M
		A. tumefaciens	SP/P20442	--TN-L--V--LI-	--YE-PT-EK-KALV--E-VKG-	NG-AWV-AQ--N	YS-S----M----M
		C. crescentus	GB/M95799	--TN-L--I--LI-	-TAS-PV-EK-KGMV-YRSSR-R	AG-AWV-AHG-D	YS--EV-------M
		B. ovis	1027504	-----L--V--LL-	--YD-PM-TK-KDLV-Y--VKG-	NG-AWVE-HG-K	YS-S----M----M
		Myxo. xanthus	1805284	---N-V-----LI-	-K-DSP-GKKAIGVS---VASSP	NG-AWVEIRG-G	YS-PEV--IV-M-M
		Hel. pylori	2072520	---K--YSI--I--	LM-NEDKAKEAEKRL-Y--VDRN	GA CAIEISG-I	Y---E---K--M-L
		The. roseum	1813674	---N--YSI--F--	---D-P----TIKLV-YQVRR-Q	NGGVAV-MGERS	Y---E---M----L
		Cyto. aquatilis	Unpubl.	--TK--ASI--F--	HT- A-TTNESKRVSY-VVKGV	QQYSTVDIDGRL	Y-A-EL--MT---M
		Fl. ferrugineum	Unpubl.	--QN--TSV--F--	-G- N--TEEI-HWSY-VAQG-	NNTVRIDIDGRL	Y---E---MV---M
		Bor. burgdorferi	SP/P28608	---N--YSI--F--	--- ---ASEIKMV-Y--EKGL	NG-ARVNISNIKKQMS-PE---AT-T-M
		Chl. trachomatis	SP/P17821	---K-LAST--FI-	-K- S--ESEIKTV-Y-VAPNS	KG-AV-D-EQ-L	Y--EE-G-Q--M-M
		Chl. pneumoniae	SP/P27542	---K-LGST--FI-	-KY S--ASEIQTV-YTVTSGS	KG-AV-E-DG-Q	Y--EE-G-Q--M-M
		Syn. sp. PCC7942 (1)	GB/D29668	--RN-FANI--FI-	--Y D-LTDESKRV-YTVRRDP	EGNVRIVCPQLSRE-A-EEVA-M--R-L
		Syn. sp. PCC7942 (2)	GB/D28551	---N-FYSV--FI-	--P D--TNELTEVAY-VD TS	GNAV-LDSSNAGKQ-A-EE---QV-R-L
		Sy. sp. PCC6803 (1)	GB/M57518	--GN-FYSV--FI-	-K- D-ITNEATEVAYSVVKDG	NGNV-LDCPAQGKQ-A-EE---QV-R-L
		Sy. sp. PCC6803 (2)	1001148	-A-N-VYSI--FI-	--W DDTVEER-RV-YNCVKGR	DDTVSVSIRGQS	Y---E---M----L
		Aph. halophytica	2804276	-A-N--YSI--FI-	--W EDTEQERNRVSYHCVPGR	DKTVDV-CWG-Q	Y---EL--M---TL
		Anabaena variabilis	2073390	--QN-F--V--YI-	--Y NELSPESKRV-YT-RKD-	VGNI-VARLN--	-AAEE---MV-K-L
		Aqu. aeolicus	2983493	D--N-VYES--FI-	-K- NE- KEAKRVSY-VVPDE	KG-AA-DIPG-L	VR-EEVG-HV-R-L
		D. proteolyticus	1813672	--QA-L-EV--FI-	--W D--KDEAARS--TVKEGP	GGSVRIE-DG-D	YA-E-V--KL-R-L
		Ther. aquaticus	GB/Y07826	-------EI--FI-	--- ----EEAKRV-Y-VVPGP	DGGVRVE-KG-L	Y--EE---M--R-L

FIG. 3. Signature sequence in Hsp70 proteins showing a specific relationship between archaebacteria (A) and gram-positive bacteria (G^+) (both monoderm prokaryotes) and the distinctness of gram-negative bacteria (G^-) (diderm prokaryotes). The large indel common in all gram-negative bacteria (referred to as the diderm insert) but absent in all monoderm prokaryotes is boxed. In the top diagram, ✪ denotes the root of the prokaryotic tree as inferred in the text. The thick arrow indicates the probable stage where this signature was introduced. The dashes in all sequence alignments show identity to the amino acids in the top line.

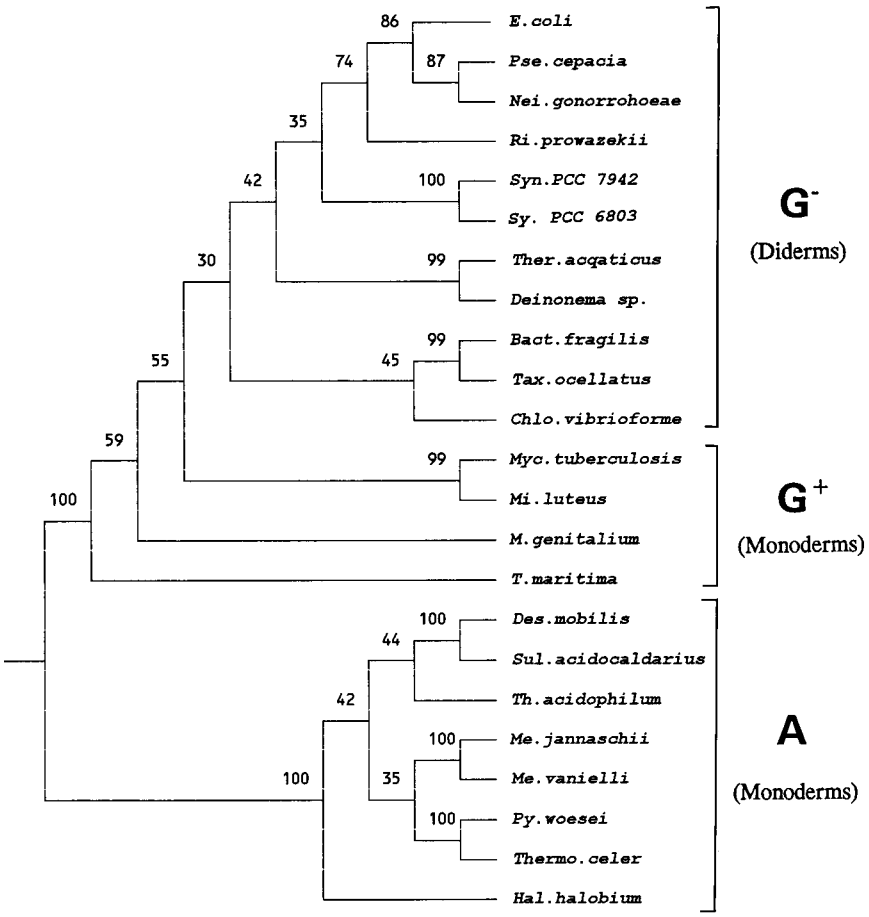


FIG. 4. A rooted neighbor-joining tree of prokaryotic organisms based on EF-1α/Tu sequences. The tree was rooted by using aligned EF-2/G sequences, which are derived from an ancient gene duplication in the common ancestor of prokaryotes (126). The tree shown is a neighbor-joining consensus tree obtained after 100 bootstraps. The bootstrap scores for various nodes are shown. The tree reveals that the root of the prokaryotes lies in between two groups of monoderm prokaryotes. A, G^+, and G^- refer to archaebacteria, gram-positive bacteria, and gram-negative bacteria, respectively.

the genomes of *M. genitalium*, *H. influenzae*, and *M. jannaschii*. The BLAST program compares a given query sequence against all other proteins and nucleic acid sequences in the databases (with the nucleic acid sequences translated in different possible frames) to identify related proteins and present them in the order from highest to lowest similarity scores. For many proteins, too few high-scoring sequences, which are suggestive of true homologs, were available to be useful for evolutionary studies. These proteins were not further considered at this stage. However, for proteins for which sufficient high-scoring sequences were identified from the major groups of prokaryotes, the sequences for various homologs were retrieved and a multiple sequence alignment was created with the CLUSTAL program (117). The sequence alignments were inspected visually for signature sequences (indels) that were shared by all members from particular taxa of prokaryotes. The indels which were not flanked by conserved regions were judged to be unreliable and were not considered as signature sequences in the present work. In cases where useful sequence signatures were observed in prokaryotic organisms, homologs from eukaryotic species were also retrieved and aligned to determine the relationship to the prokaryotes. Much of the work on the identification of signature sequences was completed by October 1997, and hence information released after this date may not be included here.

EVOLUTIONARY RELATIONSHIPS AMONG PROKARYOTES

That the ancestral organisms were prokaryotes and that the eukaryotes originated from these at a later time is a view consistent with the fossil record, which supports the existence of prokaryotic organisms as far back as 3.5 to 3.8 Ga whereas the earliest identifiable eukaryotic fossils are only about 1.8 Ga old (30a, 141, 162, 209). The Earth's geological and environmental history also supports this view (Fig. 6). There is good evidence that for the first 2.0 to 2.5 Ga, the Earth's atmosphere contained little, if any, oxygen, and hence the earliest organisms that evolved were anaerobic, whereas aerobic organisms evolved from these at a later time (132, 141, 208, 209). Since most eukaryotic organisms require oxygen for growth, it is very likely that they arose at a time when the atmospheric oxygen content was stable and relatively high (162, 208). There is thus little doubt that "All of the planet's early evolutionary history and well over 90% of life's phylogenetic diversity lie in the microbial world" (183). In view of this, the problem of understanding the evolutionary relationships among living organisms could be divided into two distinct parts. In the first part, we will examine the evolutionary relationship within the prokaryotes that pre-dated the eukaryotes. In the second part, based on our understanding of the prokaryotes, we will try to determine how

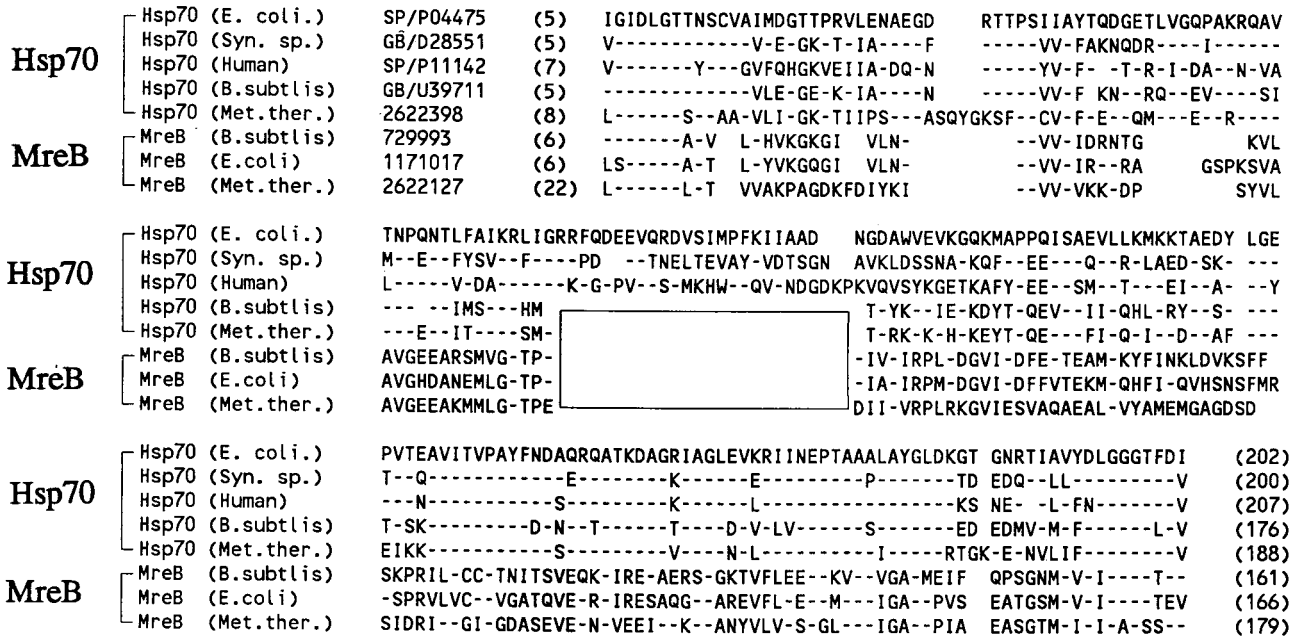


FIG. 5. Alignment of Hsp70 and MreB sequences from different groups of species showing the absence of the diderm insert in the MreB sequences. The absence of the insert in all MreB proteins, as well as Hsp70 homologs from archaebacteria and gram-positive bacteria (boxed region), provides evidence that the homologs lacking the insert are ancestral. (104, 107). The numbers at the beginning and at the end indicate the position of the sequence in individual proteins.

eukaryotic organisms are related to the prokaryotes. It should be emphasized that these two questions are completely independent. Therefore, while considering the evolutionary relationships within prokaryotes, there is no need to confound or bias the evolutionary relationships by considering sequences from various prokaryotes and eukaryotes at the same time, as has been commonly done in most earlier studies (7, 21, 49, 53, 69, 70, 81, 112, 126, 196, 198, 258, 262).

Signature Sequences Showing the Distinctness of Archaebacteria

Signature sequences consisting of distinct nucleotides that are present at particular positions in the SSU rRNA and that distinguish archaebacteria from other prokaryotes have been described by Woese (251, 253). The view that archaebacteria are distinct from other prokaryotes is also supported by signature sequences in many proteins. The elongation factor EF-1/Tu provides a well-studied example (Fig. 7a), where a 12-aa indel is present in various archaebacteria but not in any of the eubacteria including different genera of gram-positive bacteria (99, 112). Some other proteins where signature sequences unique to archaebacteria are found include ribosomal proteins L5 (Fig. 7b), S5 (Fig. 7c), and L14 (Fig. 7d). As expected from these signatures, the inference that archaebacteria are distinct from other prokaryotes is strongly supported by phylogenetic analyses based on rRNA, EF-1/Tu, and these other proteins (7, 21, 71, 87, 112, 126, 250). For all the above proteins, the

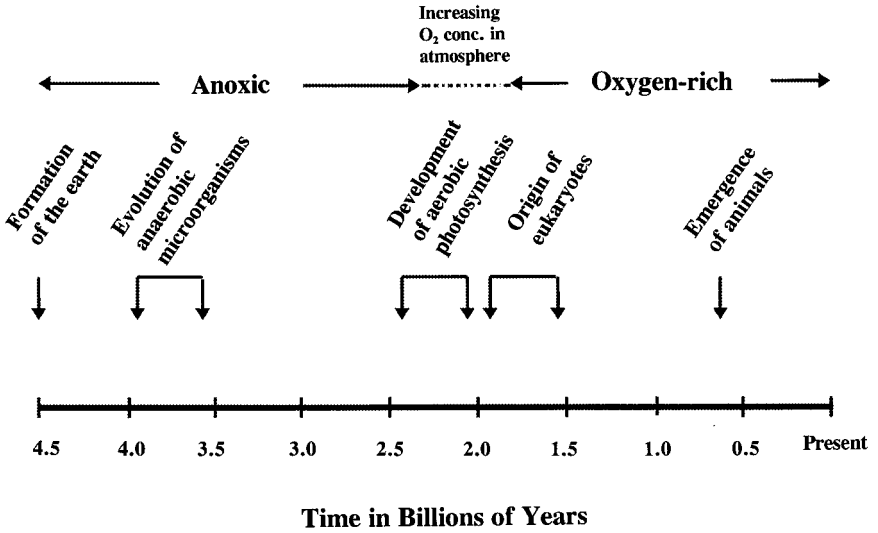
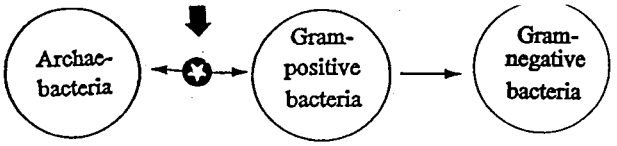


FIG. 6. Time line showing some of the main events in the history of this planet based on geological and fossil evidence (132, 141, 208, 209).



(a)

			12		71
A	Thermo.celer	P17197	IGHVDHGKSTTIGR	LLFDTANIPENI	IKKF EEMGEKGKS FKFAWVMDRLKEERERGITI
	Hal.marismortui	P16018	----------LV--	--YE-GSV--HV	-EQHK--AE----GG-E--Y---N-A------V--
	Hal. halobium	352354	----------MV--	--YE-GSV--HV	-EQHK--AE-E--GG-E--Y---N-A------V--
	Py.woesei	P26751	----------	--Y--G----Q-	---- --------- ----------R---------
	Th.acidophilum	P19486	----------LV--	--YEHGE--AH-	-EEYRK-AEQ---AT-E-------F-------V--
	Met. thermoauto.	2622158	----------LV-H	--LQAGA-A-QQ	L AE--D --RF------S------V--
	Me. jannaschii	2494244	-----A-----V--	--Y-SGA-DPQL	LE-LKR-AQ-R--AG-E--Y---N--------V--
	Archaeo. fulgidus	2649659	----------L---	--YE-GE---H-	-E-MRK-AQ----AT-E----------V--
	Me.vannielii	P07810	-----A-----V--	--L-GGA-DPQL	-VRLRK-AE----AG-E--Y---G--------V--
	Sul.acidocaldarius	P17196	----------L---	--M-RGF-D-KT	V-EAE-AAKKL--DSE-Y-FL----------V--
	Sul.solfataricus	P35021	----------LV--	--M-RGF-D-KT	V-EAE-AAKKL--ESE---FLL---------V--
	Des.mobilis	P41203	----------MT-H	I-YRLGYFD-KT	V-MIE--SKKM--ES-----LL--M--------V--
E	En.histolytica	X83684	-----S-----T-H	-IYKCGG-DQRT	-E--EK-SA-M--GS--Y---L-N--A--------
	Rh.racemosus	P14865	-----S-----T-H	-IYKCGG-DKRT	-EE-EK-AA-L--GS--Y---L-K--A--------
	S.cerevisiae	P02994	-----S-----T-H	-IYKCGG-DKRT	-E--EK-AA-L--GS--Y---L-K--A--------
	Human	P04720	-----S-----T-H	-IYKCGG-DKRT	-E--EK-AA-M--GS--Y---L-K--A--------
	Tomato	P17786	-----S-----T-H	-IYKLGG-DKRV	-ER-EK-AA-MN-RS--Y---L-K--A--------
	Di.discoiedium	P18624	-----A-----T-H	-IYKCGG-DKRV	-E-YEK-AS-M--QS--Y-----K--A--------
G^+	Myc.leprae	P30768	--------T-LTAA		-T-VLHDKFPDLNETKA-DQ I-NAP---Q-----
	Str.coelicolor	X77039	---I----T-LTAA		-T-VLHDAYPDINEASA-DQ I-KAP---Q-----
	Mi.luteus	P09953	--------T-LTAA		-S-VLYDKYPDLNEARD--T I-SAP---Q-----
	T.maritima	M27479	---I------LTAA		-T-YLSLKV LAQYIPYDQ I-KAP--KA-----
	Sp.platenesis	P13552	--------T-LTAA		-TMTLAAS- -AKAR-YDD I-AAP--KQ-----
	Bac.subtilis	P33166	--------T-LTAA		-TTVLHKKS GKGTAMAYDQ I-GAP---------
	M.genitalium	P18906	---I----T-LTAA		-CTVLSKA- TSEAK-YDE I-AAP--KA-----
G^-	E.coli	P02990	--------T-LTAA		-TTVLAKTY -GAARA-DQ I-NAP--KA-----
	Pse.cepacia	P33167	--------T-LTAA		-TTVLTKKF -GEAKAYDQ I-AAP--KA-----
	Bact.fragilis	P33165	--------T-LTAA		-TTVLAKK- LSELRS-DS I-NAP--K------
	Synechococcus sp.	416944	--------T-LTAA		-TTVLAKA- MAKARAY-D I-AAP--KA-----
	Chl.trachomatis	P26622	-------RT-LTAA		-TRTLSGD- LADFRDYSS I-NTP--KA---P-
	Bor.burgdorferi	P23125	--------T-LTAA		-SIYCSKLN -DAKAL-YED I-NAP--KA-----
	Ri.prowazekii	P48865	--------TSLTAA		-TIILAKT- -AKATAYDQ I-AAP--K------
	Chloro.auranticus	P42472	------T-LTAA		-T-VMSLK- AAQFMAYDQ I-NAP---A-----

(b)

			30		92
G^-	E. coli	132993	RVEKITLNMGVGEAIAD	KKL	LDNAAADLAAISGQKPLITKARKSVAGFKIRQGYPIGCKVTLR
	Ac. kondoi (endosymb)	710617	----------	---	----------
	H. influenzae	1173053	-I----------LT-	---	----V----------V----------
	Hel. pylori	2500236	KL---VISV-A-AHAK-	M-I	MQ-I-QTISL-A---AV----K---------E-MAV-A-----
	Bor. burgdorferi	2688401	KL---VISV-----VRN	---	--S-VLE--Q-T---AVK---K-AI--------QE--A-----
	Chl. trachomatis	132990	VLK--VIS--LA--AK-	-N-	FQAHLEE--V-------V-R-KN-I----L-E-QG--A-----
	Sy. sp. PCC6803	1652415	KLT-V-V-R-L---SQN	A-A	-ESSLTE--T-T----VV-R---AI------E-M-V-VM----
	Ther. aquaticus	243185	-L--VVI-Q-L---KE-	ARI	-EK-SKE--L-A----A--R-K--ISN--L-K-M---LR----
G^+	Mi. luteus	417671	GLV-VVV-------AK-	S-I	I-D-VT--T--T----M-------I-Q--L-E-M---THA---
	Myc. tuberculosis	1806184	T-T-VVV-------AR-	A--	ING-VN---L-T----EVRR----I-Q--L-E-M-V-VR----
	Bac. subtilis	1044976	KI---VI-----D-VQN	A-A	I-S-VEE-TF-A----VV-R-K--I---RL-E-M---A-----
	T. maritima	437935	KLV--VI---I--GSRN	YD-	IERH-NE--K-T----IV-R----ISN----K-M---L-----
	M. genitalium	1045847	KLT--VV-----D--R-	N-F	-ES-LNE-HL-T----VA---KNAISTY-L-A-QL--------
A	Hal. marismortui	132996	-I--VVVH--I-HGGR-		-A--EDI-GE-T--M-VR---KRT-GE-D--E-D---A-----
	Me. jannaschii	1710572	-I--VVV-F----SGDR		-TKG-QVIEELT----IR-R-KQTNPS-G--KKL---L-----
	Me. vannielii	132997	-IQ-V-V-F-----GDR		-TIG-KVIETLT--A-VR-L-KQTNPA-G--KKL---L-----
	Archaeo. fulgidus	2648645	VLD-VVI-I----SGER		HKK-YSL-EELVE---A--Y-KMTIKN-G--K-EA--I-----
	Sul. acidocaldarius	243187	VLD-V-V-I----SGER		-QK-YQLVQELT-V--VY--G---IRE-GV-K-A---V-A---
	Th. acidophilum	1873336	IID-VVV-I---Q-GDR		-TK--KV-EMLT-H-ATN-L-K--IRD-N--KRL---V-----
E	S.cerevisiae	914973	KI--LV--IS---SGDR		-TR-SKV-EQL---T-VQS---YT-RT-G--RNEK-AVH--V-
	Schiz. pombe	1710494	-IS-LV--ISL--SGDR		-TR--KV-EQL---T-VFS---YTIRR-G--RNEK-A-H--V-
	Dr.melanogaster	558485	HIR-LC--IC---SGDR		-TR--KV-EQLT--Q-VFS---YT-RS-G--RNEK-AVHC-V-
	Pig	971762	-IR-LC--IC---SGDR		-TRX-KV-EQLT--T-VFS---YT-RS-G--RNEK-AVHC-V-
	Human	1350658	-IR-LC--IC---SGDR		-TR--KV-EQLT--T-VFS---YT-RS-G--RNEK-AVHC-V-
	Rice	2570507	K-Q-LV--IS---SGDR		-TR-SKV-EQLI--S-VFS---YT-RS-G--RNEK-A-Y--V-

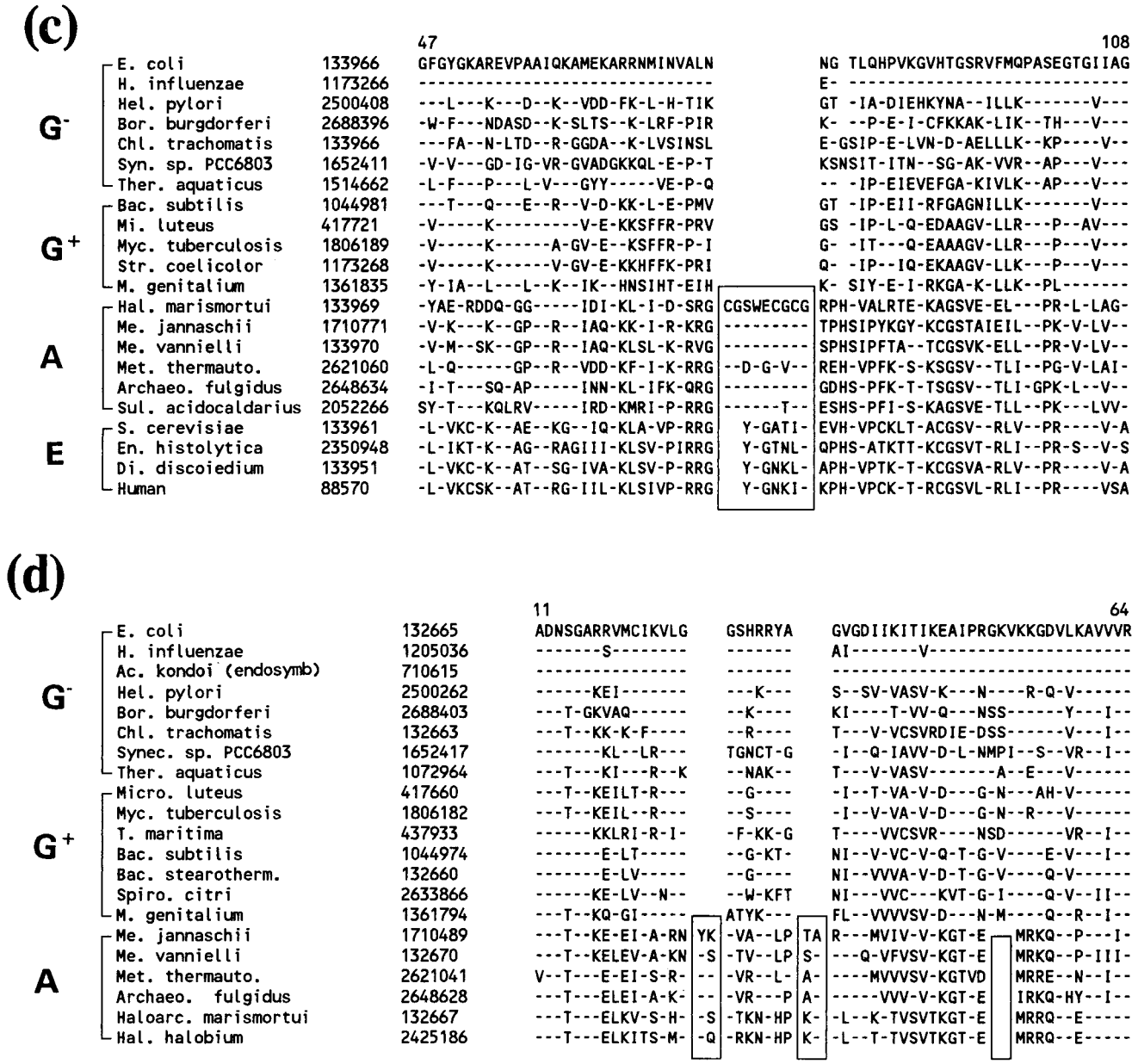


FIG. 7. Excerpts from EF-1α/Tu (a), ribosomal protein L5 (b), ribosomal protein S5 (c), and ribosomal protein L14 (d) alignments identifying signature sequences that show the distinctness of archaebacteria (A) from eubacteria (G^+ and G^-). The common indels that distinguish archaebacteria from eubacteria are boxed. (The relationship of prokaryotes to the eukaryotes [E] is considered in later figures). ✪ denotes the root of the prokaryotic tree as inferred in the text, and the thick arrow indicates the probable stage where these signatures were introduced.

identified signature sequences are present only in archaebacteria but not in any eubacteria. These results support the view that archaebacteria are monophyletic and distinct from other prokaryotes. (The question of archaebacterial monophyly and of the evolutionary relationships within archaebacteria is examined in detail below; see "Nature of the archaebacterial group and its relationship to gram-positive bacteria"). In my working model, which places the root between archaebacteria and gram-positive bacteria, these signatures were probably introduced in a common ancestor of either archaebacteria or gram-positive bacteria after separation of the two lineages (diagram in Fig. 7). In addition to these signature sequences, the distinctness of archaebacteria from eubacteria is supported by a number of other genes (21, 144, 183). These include large-subunit (LSU) rRNA (48); many genes involved in DNA replication (54), transcription (158, 197, 203), translation (47, 183), tRNA splicing (9), and in histones (197); and the Tcp-1 chaperonin (96). For a number of these genes and proteins, no eubacterial homologs or any closely related eubacterial homologs have been found (9, 21, 47, 54, 95, 96, 144, 183, 197).

Signature Sequences Distinguishing Archaebacteria and Gram-Positive Bacteria from Gram-Negative Bacteria

A specific relationship between archaebacteria and gram-positive bacteria to the exclusion of other prokaryotes is sug-

gested by a number of protein sequences. The Hsp70 protein discussed above provides the best-studied example of such sequences. As seen in Fig. 3, the Hsp70 homologs from various archaebacteria and gram-positive bacteria are distinguished from all other prokaryotic homologs by the absence of the large diderm insert in their N-terminal quadrant. The species which do not contain the insert include the methanogenic, thermoacidophilic (*Thermoplasma acidophilum*), and halophilic archaebacteria and different genera of low-G+C and high-G+C gram-positive bacteria. The *Mycoplasma* species, which lack a cell wall, and a number of other species, e.g., *Thermotoga maritima* and *Megasphaera elsdenii*, showing anomalous Gram staining (251), also lacked the diderm insert, providing strong evidence for their placement in this group. In contrast to these groups, all members of other eubacterial divisions, including the alpha, beta, gamma, delta, and epsilon subdivisions of the proteobacteria, chlamydias, spirochetes, cytophagas, flavobacteria, cyanobacteria, green nonsulfur bacteria, *Deinococcus*, *Thermus*, and *Aquifex*, which traditionally form the gram-negative group, contained this insert. The inference from this shared signature sequence that archaebacteria are specific relatives of and more closely related to gram-positive bacteria than to gram-negative bacteria is strongly supported by the detailed phylogenetic analyses based on Hsp70 sequences (56, 57, 85, 103, 104, 108). A neighbor-joining tree based on Hsp70 sequences is shown in Fig. 8. Various archaebacteria and gram-positive bacteria grouped together in 99% of the bootstraps, indicating strongly that they are evolutionarily closely related. In contrast, all of the gram-negative bacteria formed a separate clade, indicating their phylogenetic distinctness. A close relationship of archaebacteria to gram-positive bacteria and the distinctness of gram-negative bacteria are also supported by other phylogenetic methods such as maximum parsimony and maximum likelihood (85, 104, 108). Furthermore, it should be noted that the archaebacterial species in the Hsp70 tree do not form a monophyletic clade but instead show polyphyletic branching within gram-positive bacteria. The significance and possible interpretations of this observation are discussed below (see "Nature of the archaebacterial group and its relationship to gram-positive bacteria").

A close and specific relationship between archaebacteria and gram-positive bacteria is also supported by signature sequences in a number of other proteins. In the glutamine synthetase I (GS I) sequences, a conserved insert of 26 aa is present in all gram-negative bacteria but not in various archaebacteria or gram-positive bacteria (Fig. 9a) (22). Similar to the Hsp70 sequence, *T. maritima* also lacked the insert in its GS I sequence, supporting the view that it is a gram-positive bacterium. *Aquifex aeolicus*, on the other hand contained this insert, supporting its grouping with gram-negative bacteria as indicated by its Hsp70 sequence signature (Fig. 3). Phylogenetic analyses of GS I sequences again strongly support the view that archaebacteria are evolutionarily close relatives of the gram-positive bacteria and show polyphyletic branching within them (22, 85, 239). It should be mentioned that although both Hsp70 and GS I sequences show similar relationships, the presence of two different families of GS sequences in a number of different soil bacteria means that the evolutionary inferences based on GS I sequences are not as clear-cut as those based on Hsp70 (22, 239). The GS II homologs from certain soil bacteria including some gram-positive bacteria (*Streptomyces coelicolor* and *S. roseosporus*) contain the insert, which is absent in their GS I sequences (Fig. 9a). These homologs in gram-positive bacteria are likely derived by means of horizontal gene transfer from the gram-negative species (146a).

The protein glutamate-1-semialdehyde 2,1-aminomutase provides another example where a conserved indel is shared by various archaebacteria and gram-positive bacteria (Fig. 9b) but not by any of the gram-negative bacteria that have been examined.

The presence of signatures that are common to archaebacteria and gram-positive bacteria but not present in gram-negative bacteria is best explained in terms of their introduction in a common ancestor of all gram-negative bacteria (Fig. 9, top diagram). Further, based on these signatures and those distinctive for archaebacteria (Fig. 7), it is clear that gram-positive bacteria are related on the one hand to archaebacteria and on the other to gram-negative bacteria. Thus, gram-positive bacteria occupy an intermediate position between archaebacteria and gram-negative bacteria, and based on the rooting, the latter group has evolved from them.

The distinctness of gram-positive bacteria from gram-negative bacteria is also supported by signature sequences in a number of other proteins. In the highly conserved Hsp60 or GroEL protein, where sequence information is available for most of the known bacterial phyla, including different subdivisions of proteobacteria, chlamydia, spirochetes, cytophagas, flavobacteria, cyanobacteria, *Deinococcus*, *Thermus*, *Aquifex*, and different groups of gram-positive bacteria, a 1-aa insert is present in various gram-negative bacteria (Fig. 10). The species *Thermus aquaticus* and *Deinococcus proteolyticus*, which contain an outer membrane, are exceptions which are discussed below (see "Signature sequences indicating that *Deinococcus* and *Thermus* are intermediates in the transition from gram-positive to gram-negative bacteria"). Additional examples of proteins which show similar behavior to Hsp60 are also described in this later section. In phylogenetic trees based on Hsp60 sequences, the gram-negative bacteria form a monophyletic clade distinct from various gram-positive bacteria (Fig. 11) (96, 98, 246). The tree shown in Figure 11 is unrooted. However, in other studies where the Hsp60 tree was rooted with the TCP-1 protein, which is a distant Hsp60 homolog present in archaebacteria (95, 243), the low-G+C gram-positive bacteria were the deepest-branching group within the eubacteria (96, 98).

A Specific Relationship between Archaebacteria and Gram-Positive Bacteria and the Distinctness of Gram-Negative Bacteria Is Consistent with Prokaryotic Cell Structures and Other Gene Phylogenies

The presence of the indicated signatures in Hsp70, GroEL, and GS I sequences in all members of the main phyla or divisions within the gram-negative bacteria provides evidence that this group of prokaryotes is monophyletic and distinct from archaebacteria and gram-positive bacteria. This inference is in sharp contrast to that reached based on SSU rRNA sequences. The trees based on SSU rRNA generally place gram-positive bacteria between different divisions of gram-negative bacteria (55, 75, 181, 184, 250, 251). The eubacterial divisions, consisting of *Thermotogales*, green nonsulfur bacteria, deinococci, and cyanobacteria, generally show deeper branching than do gram-positive species, whereas other divisions, including proteobacteria, planctomycetes, spirochetes, chlamydiae, cytophagas, and flavobacteria, branch either lower than or in a similar position to gram-positive bacteria (35, 181, 184, 250, 251). However, most published eubacterial phylogenies based on rRNA do not give any bootstrap scores or other measures by which the confidence of these branching orders may be assessed (181, 184, 250, 251). In a few cases, where bootstrap scores are indicated, the values for most of the critical nodes leading to gram-positive bacteria are in the range of

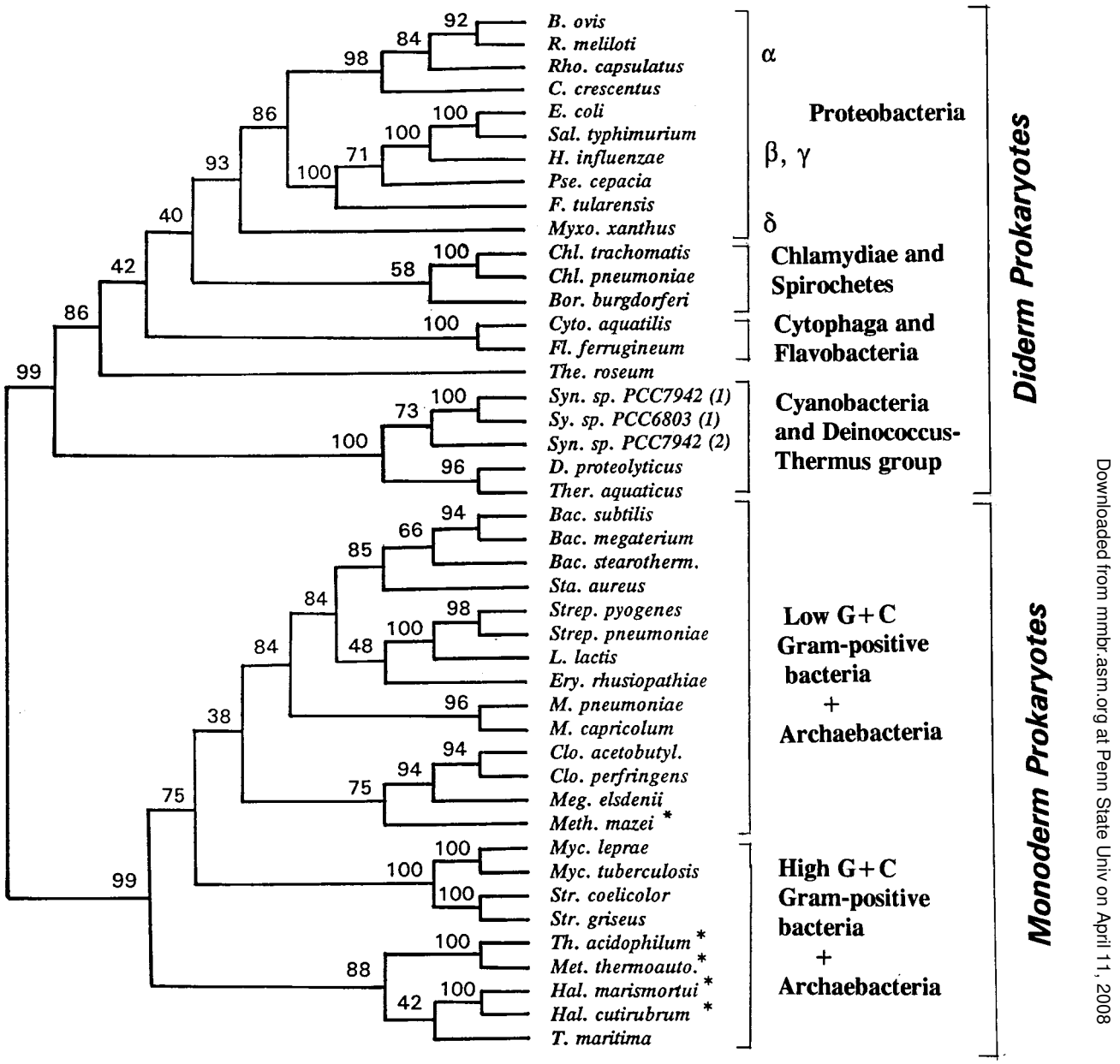
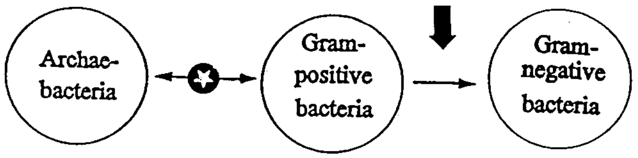


FIG. 8. Consensus neighbor-joining tree for prokaryotic organisms based on Hsp70 protein sequences. The tree, which was bootstrapped 100 times, is based on 362 aligned positions for which sequence information from all species are known. Other trees based on larger numbers of aligned characters also show similar results (see Fig. 27) (57, 103, 108). The archaebacterial species (marked with asterisks) show a polyphyletic branching within gram-positive bacteria (both monoderm prokaryotes), which is statistically strongly supported (95, 108). The gram-negative bacteria (diderm prokaryotes) form a distinct clade in 99% of the bootstraps, which is highly significant. The relationships and branching orders of some of the main divisions within eubacteria are indicated.

25 to 50%, indicating that these branching orders are unreliable (55, 75). Thus, as acknowledged by Woese (251, 252), the branching orders of major eubacterial phyla cannot be resolved based on SSU rRNA phylogenies. A number of other gene and protein phylogenies that have been previously studied, e.g., 5S rRNA (122, 211), LSU rRNA (48), EF-Tu (42, 126), EF-G (16, 42), Rho (185), aspartate aminotransferase (249), glyceraldehyde-3-phosphate dehydrogenase (114), sigma factor 70 (94), aminoacyl-tRNA synthetases (20), and RecA (55, 247), and a large number of proteins examined by Brown and Doolittle (21) similarly lacked the resolution to clarify the relationship between gram-positive bacteria and gram-negative bacteria. The inability of these phylogenies to resolve this relationship occurs in part because these genes and proteins are not highly conserved (see also other factors discussed in "Molecular phylogenies: assumptions, limitations, and pitfalls"), and for many of them only limited representation of eubacterial phyla was available.

Although most earlier gene phylogenies did not resolve the relationship between gram-positive and gram-negative bacteria, it is important to note that in the vast majority of these cases, gram-positive prokaryotes were indicated to be the closest relatives of archaebacteria. For example, in the reported phylogenies for 16S rRNA, EF-1α/Tu, EF-2/G, RNA polymerase, aminoacyl-tRNA synthetases, and various ribosomal proteins, which form the basis for defining archaebacteria as a



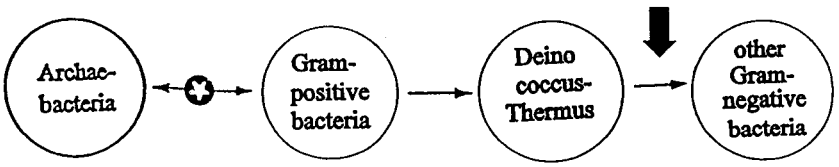
(a)

Group	Species	Accession	139	Insert	173
A	Methan. jannaschii	2118351	VGPEPEFFLLKRDPHNPHRWV		PADDGGYFDVEPLD
A	Methan. voltae	121370	--------I-- NENGK--		-G--A---EL----
A	Met. thermoauto.	2622693	--------I-D QDEDGNII		-H-C-A------V-
A	Halo. volcanii	1169928	-A-------FEE-EDGRATT-		TN-A-----LA-K-
A	Py. furiosus	462181	I------Y-F- KNGTWELE		IP-V-----ILT--
A	Py. woesi	544394	I------Y-F- KNGTWELE		IP-V-----ILT--
G+	Bac. subtilis	121359	L--------F-L-EKGEPTLE		LN-K-----LA-T-
G+	Bac. cerus	121357	L--------F-V-EKGNPTLE		LN-N-----LA-M-
G+	Clo. acetobutylicum	121362	----C----FET-ENGRATTN		TQ-KA----LA-T-
G+	Lac. delbrueckii	1169930	-F-A----F-EGKNGEETTK		VS-HSS---MASE-
G+	Sta. aureus	1134886	L--------F-L-EKGEPTLE		LN-------LA-T-
G+	T. maritima	544395	A---M---I-PINEKGEPVPE		FL-H-----LL--S
G+	Myc. tuberculosis	1707958	-H--I------PG-EDGSVP-		-V-NA----QAVH-
G-	E. coli	417057	F--------FDDIRFGSSISG	SHVAIDDIEGAWNSSTQY EGGNKGHR	--VK----P-P-V-
G-	Sal. typhimurium	1169934	F--------FDDIRFGASISG	----------------K- --------	-GVK----P-P-V-
G-	H. influenzae	1169927	F--------FDDVRF-VSMNK	ASFS-----A---TNKK- -E--NAY-	-LKK---CA-A-I-
G-	Azo. vinelandii	121356	F-------IFDEVKYKSDISG	-MFK-FSEQA---TDADF --------	-GVK----P-P-V-
G-	Az. brasilense	121355	F---A---VFDDVKFKVEMNK	VSYEF-SE--PYT-DKD- -D--L---	-GVK----P-A-V-
G-	Vib. alginolyticus	121383	I--------FDDVKFATDMSG	-FFK---V-A---TGSD- -E------	-GVK----P-A-V-
G-	Nei. gonarrhoae	121372	F-------VFDGVEFETDMHK	TRYE-TSES---A-GLHM D-Q-T---	--VK---AP-A-I-
G-	Pro. vulgaris	121375	F--------FDDIRFKNDISG	ASY--N---A---TN-K- -D------	-MVK----PLP-V-
G-	R. leguminosarum	121335	--R-A---VFDDVKYKADPYN	TGFKL-ST-LPS-DD-D- -T--L---	-RVK----P-P-V-
G-	R. meliloti	1245379	----A---VFDDVKYKADPYN	TGFKL-SS-LPS-DD-D- -T--L---	-RVK----P-P-V-
G-	Rho. capsulatus	1707964	F---A---IFDDVRYSVTPAK	VAYQ--AEAA---TDAEV -M--LA--	AGHK----P-N-V-
G-	Rho. sphaeroides	1169933	F---A---IFDDVRYSVTPAK	VAYQ--ADA----TDSE- -M--LA--	AGHK----P-N-I-
G-	Methyl. capsulatus	121369	F---N---IFDDVRWGANMSG	-FYKV-SE-AG---EKV- -D--I---	-GVK----P-P-V-
G-	Hel. pylori	2494740	F-A-N---IFDSIKIKDASNS	QYYEV-SE--E--RDRSFEN-V-F---	-GKQ---MP-P-T-
G-	Thio. ferroxidans	121382	F-- ---VFDSVTW-IDMSG	CAYKV-AE-A----GKE- -S--M---	LGVK----P-P-V-
G-	Fr. diplosiphon	417058	F---A---IFDDARFDQTANS	GYYYV-SV--R---GKD --P-LAYK	-RFKE---P-A-T-
G-	Aqu. aeolicus	2982851	F---A---IFDSVEFGTAANY	AFWRV-SE--W--R -VPSS-YK	IPHKR---PAP-V-
G-	Sy. sp PCC6803	1652131	F---A----FDDIRFGQTENS	-YYFA-SV--R--TGREE ----L-YK	-GYKQ---P-A-T-
G-	Syn. sp. PCC7002	121381	F---A---VFDDVRFDQTENK	GFYYV-SV--R---GRKE P---LA-K	-GYKQ---P-P-T-
G+ (II)	Str. coelicolor	121379	F---A--YVFDSVRFATRENE	-FYH--SEA----TGA LED-R-YK	VRYK----P-P-V-
G+ (II)	Str. rosesporus	2494750	F---A--YVFDNVRFQTSANE	-FYH--SEA----TGA VEN-R-YK	VRYK----PAP-V-

(b)

Group	Species	Accession	294	Box	360
G-	E. coli	121655	YQAGTLSGNPIAMAAGFACLNEVAQPG	V	HETLDELTTRLAEGLLEAAEEAGIPLVVNHVGGMFGIFF
G-	Sal. typhimurium	121657	---------------------------	I	---------------C---Q------------------
G-	Pse. aeruginosa	1346186	----------L-----LTT-RLISR--	F	-DE-TAY---MLD--QQR-DA----F-TTQA-----LY-
G-	Xan. campestris	544428	----------V-----L-M-EL-QE--	F	-TR-S-A-SM-C---ED--RA---AVTT-Q------L--
G-	Hel. pylori	2492857	----------L-VC--LSA-YKIKRDK	T	LY-RLDALAIRLTQG-QKSAQNYNIALETLNM-SMFG--
G-	Aqu. aeolicus	2983378	------A---L--VS-SET-KDLRDKE	P	YKE-E-KMEK--R-VKDILT-K--QHTI-K--S-MTV--
G-	Syn. sp. PCC6301	1170032	----------L--T--IKT-ELLR---	T	Y-Y--QI-K--SD---AI-Q-T-HAACGGQ-S----F--
G-	Sy. sp. PCC6803	1001558	----------L--T--IKT-EILQK--	S	Y-Y--KI-K--VD---A--QD--HEVCGGSISA------
E	Sol lycopersicum	642911	----------L--T--IHT-KRLQGQ-	T	Y-H--KI-AE-TQ-I-D-GKKT-HAMCGGSIR----F--
E	Nicotian tabacum	100332	----------L--T--IHT-KRLQG--	T	YDY--KI-GE-TQ-I-D-GKKT-HAMCGGYIR----F--
E	Ara. thaliana	498914	----------L--T--IHT-KRLS---	T	Y-Y--KI-KE-TN-I---GKKT-HAMCGGYIS----F--
E	Glycine max	1170031	----------L--T--IET-QRIKE--	T	Y-Y--KI-GE-V--II--GKR--HAICGG-IR----F--
G+	Bac. subtilis	399784	----------L--T--LET-KQLTPES		YKNFIKKGD--E--ISKT-GAH---HTF-RA-S-I-F--
G+	Sta. aureus	2589184	----------L--TS-YET-SQLTPET		Y-YFNM-GDI-ED--KRVFAKHNV-IT--RA-S-I-Y-L
G+	Prop. shermanii	544427	----------A-C---L-T-ALMDDAA		YSR--ATAD-VSAMADA-L-S--V-HRI-K-SNL-SV-L
G+	Myc. tuberculosis	2113986	----------V-V---L-T-RAADDAV		YTA--ANAD---GL-S--LTD-VV-HQISRA-N-LSV--
G+	Myc.leprae	1170030	----------V-----L-T-RAA-DAV		YA---RNAD--VAM-S--LTD--V-HQIPRA-N--SV--
A	Me. janaschii	1591312	-----FN----SIT--I-T-KQLDDRF		YKETARTAKI--DT-R-L-DKHN-KAK-YNIAS--Q-Y-
A	Met. thermauto	15712676	-----FN---VSVT--RET-RLLDGRM		YSD-ERKGST-RA--RDLLSDLDLEYQ-TGPAS--QLY-
A	Archaeo. fulgidus	2649338	-----F----LSLT--Y-TVKFMEEN-	-	I-KVNS--EK-VS-IADVL-DKKAECE-GSLAS--C-Y-

FIG. 9. Signature sequence (boxed insert) in GS I (a) and glutamate-1-semialdehyde 2,1-aminomutase (b), showing the relatedness of archaebacterial (A) homologs to gram-positive (G^+) bacteria and the distinctness of gram-negative (G^-) bacteria. The top diagram indicates the suggested interpretation that these signature, as well as the large diderm insert in Hsp70 protein (Fig. 3), were introduced into a common ancestor of G^- bacteria. (a) G^+ (II) identifies sequences from some of the GS II family of proteins (22, 205). E, eukaryotes.



			144		178
G⁻	E. coli	P06139	IAQVGTISA	T	SDETVGKLIAEAMDKVGKEGVITVE
	Sal. typhimurium	P48217	---------	N	-------------------------
	Ch. vinosum	P31293	---------	N	--DSI-TI-----E----------
	Yer. enterocolitica	X68526	---------	N	--S---E---Q--E----------
	H. ducreyi	P31294	-E-------	N	----------Q--E----------
	Pse. aeruginosa	P30718	---------	N	---SI-QI-----E----------
	Le. pneumophila	P26878	---------	N	---AI-AI-----E----------
	Am. proteus(e)	P26004	---------	N	--QAI-SI-----E----------
	Ac. pisum(e)	P25750	-T-------	N	A--K--S------E---ND------
	Cox. burnetii	P19421	---------	N	--KSI-DI-----E----------
	Nei. flavescens	Z22955	-----S---	N	---Q--AI-----E----------
	Nei. gonorrhoeae	Z23009	-----S---	N	---Q--AI-----E----------
	R. meliloti (1)	M94192	V--------	N	GEKQI-LD-----Q---N-------
	R. meliloti (2)	M94190	----A----	N	G-AEI-RYL----E---N-------
	R. meliloti (3)	M94191	---------	N	G-TEI-RYL----E---N-------
	A. tumefaciens	P30779	V--------	N	GERQI-LD-----QR--N-------
	Ri. tsutsugamushi	P16625	----A-V-S	N	G-REI-EK--N--KQ--Q-------
	Eh. chaffeensis	L10917	----A----	N	G-KNI-SK--QCVQE---D------
	Z. mobilis	L11654	V----I---	N	G--E--RR-----E----------
	Bar. bacilliformis	P35635	----A----	N	GA-DI--M--D--E---N-------
	Hel. pylori	X73840	-T--A----	N	--HNI-----D--E----D------
	Po. gingivalis	D17398	-EH-AK---	N	G--NI-S------R--K--------
	Cyto. aquatilis	unpubl.	-K-IAS---	N	N--VI-E---T-FA------- ---
	Fl. ferrugineum	unpubl.	-Q--A----	N	N---I-------FA----------
	Chl. trachomatis	P17203	----A----	N	N-AEI-N------E----N-S----
	Bor. burgodferi	P27575	----AS---	N	N-SYI-EK----------D------
	Lep. interrogans	L14682	--N-AS---	N	N-N-I-N---D-------D------
	Tre. pallidum	S33213	V-H-ASV--	N	N-KEI-RIL-S-IE---ND---D-D
	Aqu. aeolicus	2984379	-E--A----	N	N-PEI--I--D--EE---D------
	Sy. sp. PCC6803	P22034	----A-V-S	G	TNPE--AM--D-----T-D------
	Syn.sp. PCC7942	P22879	-----A---	G	N-FE--QM--D----------SL-
Deinococcus and Thermus	Ther. aquaticus	P45746	-EE-A----		N-PE------D--E------I----
	D. proteolyticus	unpubl.	-KK-AG---		N-PQ--EE--S----------I-
G⁺	Clo. perfringens	P26821	--R-AA---		A--KI-----D--E---N-------
	Clo. thermocellum	P48212	--R-AS---		N--VI-E---D--E--TND------
	Thermoph. bact.(PS-3)	P26209	----AA---		A--E--Q------ER--ND----L-
	Bac. subtilis (1)	P28598	----AA---		A--E--S------ER--ND----I-
	Bac. subtilis (2)	Q07201	----AA---		A--E--S------ER--ND----I-
	Str. albus (1)	Q00767	--A-AAL--		Q-KQ--E----------D------
	Str. albus (2)	Q00768	--STAS---		A-TQI-E-----------------
	Str. coelicolor	P40171	--A-AAL--		Q-QQ--E----------D------
	Sta. epidermidis	P48218	-----A---		A--EI-RY-S-------ND----I-
	M. genitalium	P47632	-E--AA--S		GSKEI-----Q--AL---N----TD
	Bac. stearothermophilus	L10132	----AA---		A--E--Q------ER--ND----L-
	T. maritima	unpubl.	--H-AA---		NSAEI-E----------ED------
	L. lactis	X71132	----A-V-S		RS-K--EY-SD--ER--SD----I-
	Myc. leprae (1)	P09239	--ATAA---		G-QSI-D----------N-------
	Myc. leprae (2)	S25181	-T--A-V-S		R--QI-A-VG-G-N---TD--VS--
	Myc. tuberculosis (1)	P06806	--ATAA---		G-QSI-D----------N-------
	Myc. tuberculosis (2)	X60350	----A-V-S		R--QI-D-VG---S---HD--VS--

FIG. 10. Excerpt from the GroEL (or Hsp60) protein sequence alignment showing a 1-aa insertion (boxed) that is shared by most divisions of G^- bacteria but absent from all G^+ bacteria. The absence of this insert in *Thermus aquaticus* and *Deinococcus proteolyticus*, which are diderm prokaryotes that contain thick cell walls, indicates that this insert was introduced into an ancestral gram-negative lineage after the branching of the *Deinococcus-Thermus* group (thick arrow in the top diagram).

unique domain, the species *Thermotoga maritima*, which is now known to be gram positive, shows the closest relationship to archaebacteria (7, 20, 21, 112, 250). Brown and Doolittle (21) recently reported phylogenies based on 66 protein sequences for which sequence information was available from archaebacteria, eubacteria, and eukaryotes. They tried to determine which of the three possible relationships among these groups (i.e., an archaebacterial-bacterial clade, an archaebacterial-eu-

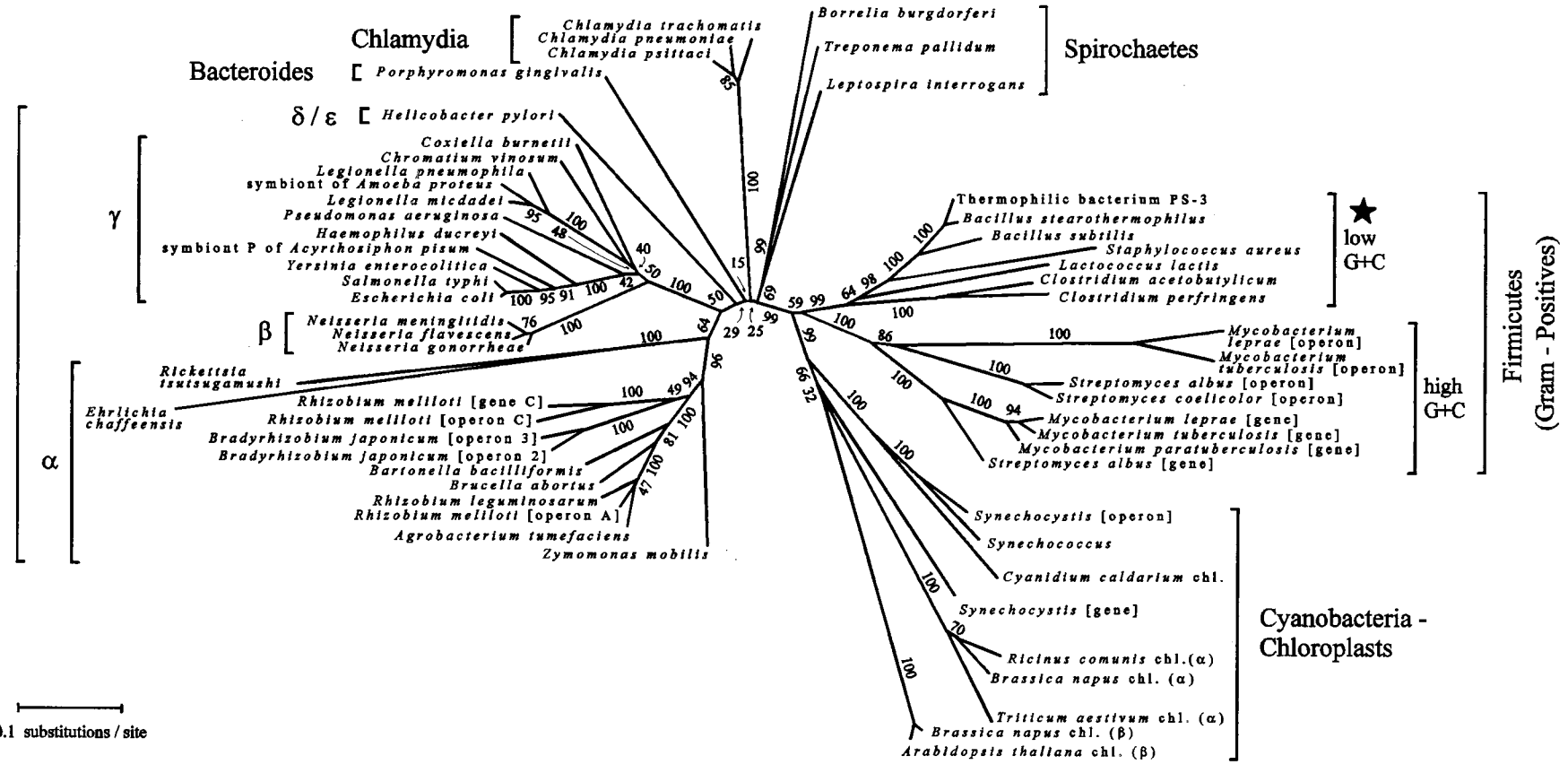


FIG. 11. Evolutionary relationships between eubacterial species and groups based on the GroEL (Hsp60) sequences. The tree shown is a consensus neighbor-joining distance tree obtained after 100 bootstraps. The distinct branching of low-G+C and high-G+C gram-positive bacteria and their close relationship to the cyanobacteria should be noted. The branching order of other prokaryotic groups in the GroEL tree is very similar to that observed for Hsp70 sequences (Fig. 8). Similar results with GroEL/Hsp60 sequences have been reported in other studies (96, 98). Although the tree shown here is unrooted, in other studies (96, 98) where the Hsp60 tree was rooted with the distantly related TCP-1 protein from archaebacteria (243), the low-G+C gram-positive bacteria were found to be the deepest-branching group within eubacteria (marked with ★). Reproduced from reference 246 with permission of the publisher.

TABLE 1. Protein phylogenies where gram-positive bacteria are indicated as the closest prokaryotic relatives of archaebacteria[a]

Protein	Fig.	Protein	Fig.
Gyrase B	6A	*trpA*	17A
Photolyase	6C	*trpB*	17B[b]
EF-Tu	8A	*trpC*	17C
EF-G/2	8B	*trpD*	17D
Isoleucyl-tRNA synthetase	9B	*hisC*	18C[b]
Tryptophanyl-tRNA synthetase	9C	*hisD*	18D
Tyrosyl-tRNA synthetase	9D	*hisF*	18E[b]
Ribosomal protein L2	10A	*hisH*	18G[b]
Ribosomal protein L11	11B	*his1E*	18H
Ribosomal protein L14	11C	IMP dehydrogenase	19A
Ribosomal protein L15	11D	FGAM synthetase	19B
Ribosomal protein L22	11E	Glutamyl-tRNA reductase	19C
Ribosomal protein L23	11F	*trpG*	17F
Ribosomal protein L30	12A	SecY	19E
Ribosomal protein S5	12B	FeMn SOD	20A[b]
Ribosomal protein S9	12E	Hsp60/Tcp-1	20B
Ribosomal protein S10	12F	Glutamine synthetase	16B
Ribosomal protein S15	13C	Glutamate dehydrogenase II	15C[b]
Ribosomal protein S19	13E	Argininosuccinate synthase	15D
Enolase	13F	Aspartate aminotransferase	16A
Acetyl-CoA synthetase	15A	Histidinol-P-aminotransferase	16A
Citrate synthase	15B	*hisG*	18F[c]
Hsp70	5[b]	RNA polymerase A subunit	7a[c]

[a] This table is based on the phylogenetic trees published by Brown and Doolittle (21). The relationships indicated are observed upon excluding the eukaryotic homologs and knowing that *T. maritima* is a gram-positive bacterium. Figure numbers in this table refer to figures in reference 21. CA, coenzyme A; SOD, superoxide dismutase.

[b] Polyphyletic branching of archaebacteria within gram-positive bacteria.

[c] For these genes, no sequences for gram-positive bacteria were included in the analyses.

karyote clade, or a bacterial-eukaryote clade) was supported by different protein phylogenies. As pointed out above, it is confounding the problem to consider the evolutionary relationships between prokaryotes and eukaryotes, as was done in this study, in the absence of a good understanding of the phylogeny of prokaryotes. However, if one examines the phylogenetic trees reported in this review (21) and asks which group of prokaryotes are the closest relatives of archaebacteria, then for more than two-thirds of the genes studied, *T. maritima* or another gram-positive bacterium was found to be the closest relative of archaebacteria (Table 1). A closer relation of archaebacteria to gram-positive bacteria has been acknowledged by these authors: "In phylogenies supporting an AB (archaebacteria-bacteria) grouping, the archael branches are often among those of the gram-positive bacteria" (21). Thus, a specific relationship of archaebacteria to gram-positive bacteria is not restricted to a few proteins but is generally observed for the majority of the gene and protein sequences (21).

What is the significance of the observed close relationship between archaebacteria and gram-positive bacteria on the one hand and the distinctness of gram-negative bacteria on the other? The answer becomes strikingly clear when the cell structures of the prokaryotes are considered (228, 241). As discussed above, based upon their cell structures, the prokaryotic organisms can be divided into two major groups—those bounded by a single membrane (termed monoderms) and those containing inner and outer membranes (termed diderms) that define the periplasmic compartment (Fig. 12). All archaebacteria and gram-positive bacteria belong to the first group. Some species which lack a cell wall (e.g., *Mycoplasma* and *Thermoplasma* species) or show gram-negative staining due to other unusual characteristics (e.g., *Megasphaera* and *Thermotoga* species) are also bounded by a single membrane. The signature sequences and phylogenies based on Hsp70 and other highly conserved proteins thus distinguish and separate all monoderm prokaryotes from the diderm prokaryotes. These results are thus in accordance with the most striking and fundamental structural difference in the organization of prokaryotes (Fig. 12, top). In addition to the presence of an outer membrane which defines the periplasmic compartment, the gram-negative bacteria differ from the gram-positive bacteria in several other respects including thickness of the cell wall, flagellar structure, and general response to the environment (124, 180a, 192, 206, 220, 241, 260). According to Tipper and Wright (241): "The Gram-negative cell has a fundamentally different strategy toward the external environment than the Gram-positive cell. In the Gram-negative cells a membrane is present, external to the peptidoglycan layer, that acts as a permeability barrier between the external environment and the cytoplasmic membrane. It is an essential component of all Gram-negative cells and apparently cannot be dispensed with, even under laboratory conditions." Thus, the inferences derived from molecular sequence data are in accordance with and strongly vindicated by the morphological characteristics of the prokaryotes.

In contrast to these results, which unite both molecular sequence data and cell structure characteristics, gram-negative bacteria (diderm prokaryotes) are not recognized as a distinct taxon in the three-domain proposal. In the three-domain proposal, while one group of monoderm prokaryotes (i.e., archaebacteria) form one domain, other domain is suggested to contain a polyphyletic branching of different monoderm and diderm prokaryotic phyla (Fig. 1a and 12, bottom) (181, 184, 250–252, 258).

Signature Sequence Distinguishing between Low-G+C and High-G+C Gram-Positive Bacteria and Pointing to a Specific Relationship of the Latter Group to the Gram-Negative Bacteria

The gram-positive bacteria are traditionally divided into two groups: the high-G+C group and the low-G+C group (6, 14,

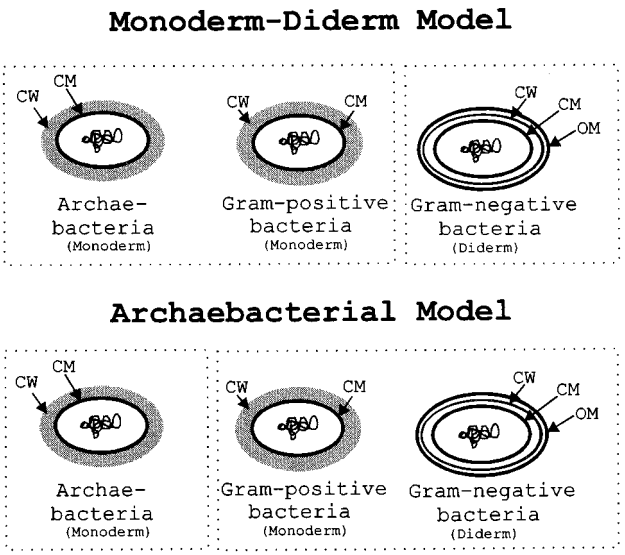


FIG. 12. Evolutionary relationships within prokaryotes as indicated by the monoderm-diderm model (top) versus the currently popular archaebacterial model (bottom). It should be noted that the latter model does not recognize diderm prokaryotes as a distinct taxon and that in phylogenetic trees based on 16S rRNA the gram-positive (monoderms) and gram-negative (diderms) bacteria show polyphyletic branching within each other (183, 250, 251, 258). Abbreviations: CM, cytoplasmic membrane; OM, outer membrane; CW, cell wall. Reproduced from reference 101 with permission of the publisher.

184, 192, 228, 229, 250). While the phylogenies based on some gene sequences, i.e., Hsp 70, GroEL (Hsp60), and sigma 70, show that these two groups are distinct from each other (94, 96, 99, 103, 108, 246), the relationship between these two subdivisions of prokaryotes is not resolved in a number of other phylogenies, including those based on SSU rRNA, LSU rRNA, RecA, EF-Tu, and EF-G (7, 16, 48, 55, 131, 184, 250, 251). Hence, the question whether these two groups are phylogenetically distinct is unclear. The signature sequences in proteins again provide important insight in this regard. In the ribosomal S12 protein, a 13-aa deletion is present in a highly conserved region in various members of the high-G+C gram-positive bacteria as well as gram-negative prokaryotes but not in any of the low-G+C gram-positive bacteria examined (Fig. 13a). Although this sequence region is not highly conserved between archaebacteria and bacteria, it is quite clear from the alignment that this deletion is also not present in any of the archaebacterial homologs. Another example of a protein showing a similar signature sequence is provided by dihydroorotate dehydrogenase, where a 2-aa insert in a conserved region is found in various gram-negative bacteria and high-G+C gram-positive bacteria examined but not in any of the low-G+C gram-positive bacteria or archaebacterial homologs. The signature sequences in these two proteins provide evidence that members of the high-G+C and the low-G+C gram-positive bacteria are phylogenetically distinct from each other. Furthermore, based on the results presented above, which suggest that archaebacteria and gram-positive bacteria are ancestral groups and that gram-negative bacteria are derived from them, the presence of these shared signatures in various high-G+C gram-positive bacteria as well as different gram-negative bacteria is strongly indicative that these two groups of prokaryotes are specifically related to each other and they had a common ancestor exclusive of the low-G+C gram-positive bacteria (Fig. 13). As shown in Fig. 13, these signature sequences were probably introduced into the main stem of the tree leading to the high-G+C gram-positive group as well as the gram-negative bacteria. These results also provide evidence that among the gram-positive bacteria, the low-G+C group is ancestral.

Additional signature sequences which appear to be specific for only the low-G+C gram-positive and the high-G+C gram-positive groups have been identified. Figure 14 shows a 2-aa insert in pyruvate kinase that seems specific for only the low-G+C gram-positive group but is not found in any of the other prokaryotic homologs. This insert, as shown, was probably introduced into the branch leading to the low-G+C gram-positive bacteria (Fig. 14). We have also come across a signature sequence in gyrase A that appears to be specific for the high-G+C gram-positive group (Fig. 15). This signature was probably introduced into the branch leading to the high-G+C gram-positive group.

Signature sequences in the above proteins also provide insights into the placement of *T. maritima* within the gram-positive group. Although *T. maritima* is clearly a Gram-positive bacterium based on signature sequences in Hsp70 (Fig. 3) and GS I (Fig. 9a), signature sequences in pyruvate kinase (Fig. 14) and gyrase A (Fig. 15) show that it does not contain the signatures that appear distinctive for either the low-G+C or high-G+C gram-positive groups. These observations indicate that *T. maritima* probably evolved from the main stem independently of the typical low-G+C and high-G+C gram-positive groups. Signature sequences in Hsp70 (Fig. 3), ribosomal S12 (Fig. 13a), and dihydroorotate dehydrogenase (Fig. 13b) suggest that this branching took place from the common ancestor of high-G+C gram-positive bacteria and gram-negative bacteria, before the evolution of gram-negative bacteria.

Signature Sequences Indicating that *Deinococcus* and *Thermus* Are Intermediates in the Transition from Gram-Positive to Gram-Negative Bacteria

The members of the genera *Deinococcus* and *Thermus* represent an interesting group of prokaryotes whose classification and evolutionary position have presented problems by both traditional and molecular criteria (19, 39, 176). As noted by Murray (176) in *Bergey's Manual*, the members of the genus *Deinococcus* show a positive Gram-staining reaction and possess a peptidoglycan component of cell wall with a thickness similar to that of the gram-positive bacteria. Accordingly, *Deinococcus* species are recorded as gram positive. However, Murray (176) also emphasized that on biochemical and structural grounds, the *Deinococcus* species are more akin to gram-negative bacteria than to gram-positive bacteria. For example, these bacteria have a fatty acid profile that is similar to the gram-negative bacteria rather than to the gram-positive bacteria (19, 39, 176). Of greater significance is the presence of an outer cell membrane in the *Deinococcus-Thermus* group (157, 176), which is a unique and defining characteristic of all gram-negative species. Phylogenies based on 16S rRNA place *Deinococcus* and *Thermus* species in a separate lineage branching below the *Thermotogales* and in a similar position to cyanobacteria and green nonsulfur bacteria (115, 181, 250, 251).

The protein phylogenies and signature sequences provide important information about the phylogenetic position of *Deinococcus* and *Thermus* within prokaryotes. Interestingly, similar to the phenotypic characteristics, the sequence data for different proteins indicate that these organisms can be grouped with either gram-positive or gram-negative bacteria. For example, the presence of the large insert in their Hsp70 protein (Fig. 3), which is a characteristic of all gram-negative bacteria, indicates that these organisms should be classified as gram-negative bacteria. This inference is in accordance with the pres-

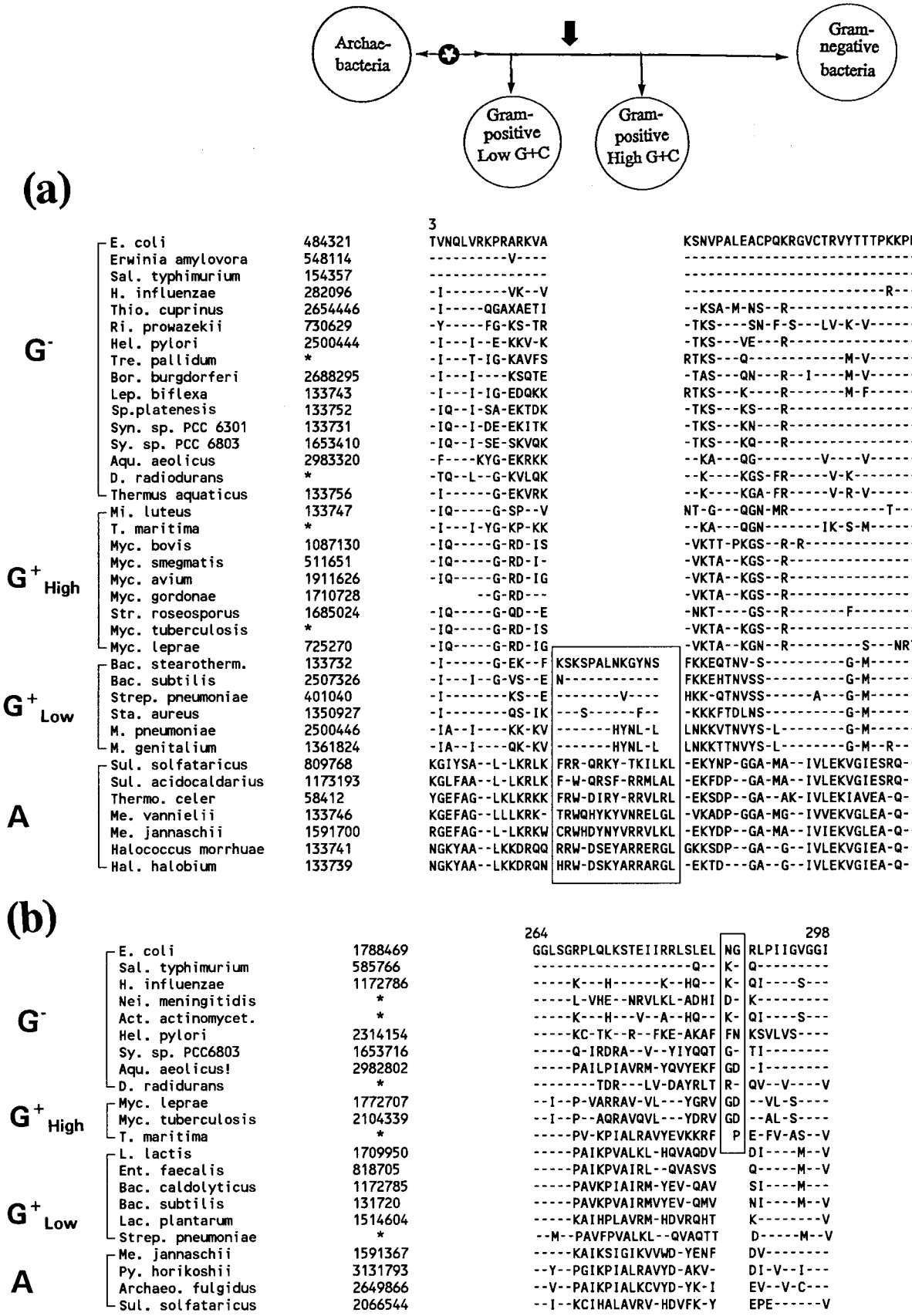
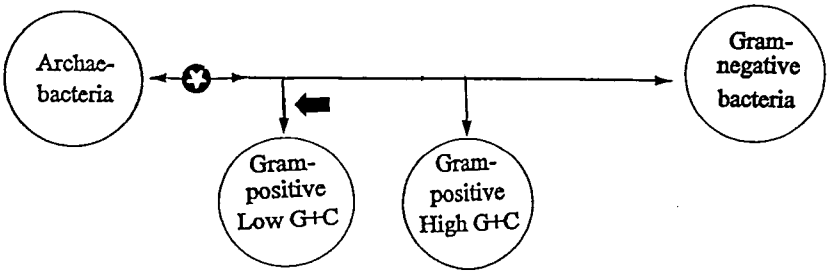


FIG. 13. Signature sequence in ribosomal S12 protein (a) and dihydroorotate dehydrogenase (b), distinguishing archaebacteria (A) and the low-G+C gram-positive bacteria from the high-G+C gram-positive group (G^+) and gram-negative bacteria (G^-). These signatures (boxed) provide evidence that the gram-negative bacteria are specifically related to the high-G+C gram-positive group. The asterisks in this and all subsequent alignments identify sequences retrieved from the National Center for Biotechnology Information unfinished microbial genomes database.



			118		178
G⁻	E. coli	125603	PADVVPGDILLLDDGRVQLKVLEVQGMK		VFTEVTVGGPLSNNKGINKLGGGLSAEALTEKD
	H. influenzae	1170698	-Q-------------------STD-A-		----------D------
	Sal. typhimurium	1526982	TS-LSV-NTV-V---LIGME-TAIE-N-		CICK-LNN-D-GE---V-LP-VSIALP--A---
	A. vitis	984370	F-A-K---D--I-----RVRA-G-SDEF		IDAK-I-A--I--R--V-LP-TV-DISP--P--
	Methylobac. extorquens	1907336	LSALE-SHGI-I---KLR-I-T--SEGR		AV-R-E---RI--R--VSLPHTA-PVP-M----
	Bor. burgdorferi	2688255	VKE-PQ-SKV-----ELEMT-VAKLPDR		LIC-IKND-QIK-K-S--TP-IS-KLQSV----
	Chl. trachomatis	1791247	-RERAPV-I---YI-AV-VNA-EHM		-EI-FQNS-EIKS--SLSIKDIDVALPFM----
	Sy. sp. PCC6803	1208543	ATEAKV-ERI-----LLEM--VSI-DPE		-IC--VT--I-KSR--V-LP-LV-TLPSM-T--
	D. radiodurans	*	AG--T--MT------NMS-R-DH-R-ND		IQ-T-LI--T-K------VPEAD-TVP--S---
G⁺ High	Cor. glutamicum	598097	AK-AK---R--V---K-G-VCVS-E-ND		-IC--VE---V-----VSLP-MDI-VP--S---
	T. maritima	*	-K--KK--TI--S--EIV-E-I-TTDTE		-K-V-K---KITHRR-V-VPTAD--V-SI-DR-
	Myc. tuberculosis	3122312	AQ-A-A--RV-V---K-A-V-DA-E-DD		-VCT-VE---V-D----SLP-MNVT-P--S---
	Myc. intracellulare	1750255	AE-AAV--RV-V---K-C-V-DGIE-DD		-ICT-VE---V------SLP-MNV--P--S---
G⁺ Low	Bac. psychrophilus	1041097	IE--NE-SVI-----LI--E-TGKDVAR	GL	IH-LIINS-S------V-IP-VSVQLPGM----
	Bac. stearothermophilus	585371	ID--SV-AKI-----LIS-E-NA-DKQA	GE	IV-T-LN--V-K-K--V-VP-VKVNLPGI----
	Bac. licheniformis	1041099	IH--SV-STI-----L-G-E-TDINKD-	RE	IV-K-MNS-T-K-K--V-VP-VSVNLPGI----
	Bac. subtilis	2293265	VH--EQ-STI-----LIG-E--D-DAA-	RE	IK-K-LNN-T-K-K--V-VP-VSVNLPGI----
	Strep. pneumoniae	*	YD--EV-RQV-V---KLG-R-VAKDDAT	RE	FEV--END-IIAKQ--V-IPNTKIPFP--A-R-
	Strep. pyogene	*	YDE-EV-HTI-I---KLG---IDKDIAT	RQ	FIV--END-IIAKQ--V-IPNTKIPFP--A-R-
	M. pneumoniae	2497533	VN--KV-QKI-V---KLS-V-KRIDTKN	NQ	-ICVAQNDHTIFTK-RL-LPNADY-IPF-SA--
	Lac. delbrueckii	1154865	FD-THV-GTV-I---A-G-TIKAKDEE-	RE	LVC-AQNT-VIGSK--V-AP-VEIRLPGI----
	Unidentified bacterium	155435	FD--QV-GQV-F---LLGTT---KDVAN	RE	LVVR-DND-I-GSR--V-AP-VSINLPGI----
	L. lactis	585372	FD--EI-QTI-I---KLG-SLTGKDAAT	RE	FEV-AQND-VIGKQ--V-IPNTKIPFP--A-R-
	M. genitalium	1045902	VN--NI-QKI-V---KLT-V-TR-DKQH	NQ	-ICVAKNDHTVFTK-RL-LPNAQY-IPF-S---
A	Thermo. litoralis	1016357	-KL-SK--TIY-S--YIM-R-E--RENE		-ECV-VN--I-FSH----IPKAN-PI--I-PR-
	Me. janaschii	1590885	IDTIEE-HFI-IN--KIK-R-V-KTDK		IIAV-E---EIKEGM-V-LPDTRIELPIID-TA

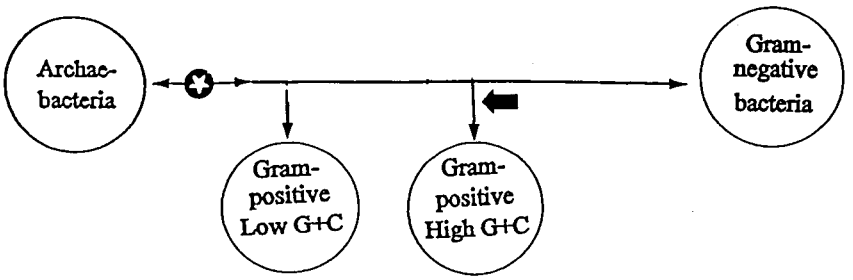
FIG. 14. Signature sequence (boxed) in pyruvate kinase which appears specific for the low-G+C gram-positive group. This signature was probably introduced in the branch leading to this particular group.

ence of an outer cell membrane in these species. In contrast, signature sequences in the Hsp60 protein (Fig. 10) indicate that *Deinococcus* and *Thermus* lack the 1-aa insert common to various other gram-negative bacteria and thus are similar to gram-positive bacteria. Signature sequences in two additional proteins, acetolactate synthase (Fig. 16a) and asparginyl-tRNA synthetase (Fig. 16b), also indicate a specific relationship of these species to the gram-positive bacteria. For asparginyl-tRNA synthetase, the absence of the insert in the proteobacterium *Helicobacter pylori* is surprising and may represent a case of horizontal gene transfer.

The above results showing a grouping of the *Deinococcus-Thermus* genera with either gram-positive or gram-negative bacteria, based on signature sequences in different proteins, are not conflicting but, instead, suggest that the members of these genera are probably derivatives of intermediates in the transition from the gram-positive to the gram-negative group of prokaryotes and hence possess some characteristics of each group. The presence of an outer cell membrane in these organisms, together with a thick cell wall in the case of *Deinococeaceae*, indicates that in the evolution of gram-negative bacteria from a high-G+C gram-positive ancestor, the outer membrane developed first before the changes in the cell wall took place. It is of interest in this context that in several mycobacterial species (high-G+C gram-positive bacteria), the membrane lipids are arranged in a highly ordered form, which may represent an early stage in the development of outer cell membrane (18a, 180b). The signature sequences in different proteins thus provide molecular markers that correlate with the phenotypic changes in the cell structure. Thus changes in Hsp70 (or GS I) which correlate with the development of the outer membrane (Fig. 3) took place in the common ancestor of gram-negative bacteria before changes in the other proteins (Hsp60, acetolactate synthase, and asparginyl-tRNA synthetase) that show correlation with the changes in the cell wall and other properties of the cells. Thus, the molecular and phenotypic characteristics of these organisms are in good agreement and point to a unique phylogenetic position of the *Deinococcus-Thermus* group as representing evolutionary intermediates.

Phylogenetic Placement of Cyanobacteria and Their Close Evolutionary Relationship to the *Deinococcus-Thermus* Group

The signature sequences discussed thus far have allowed us to reconstruct the evolutionary history from monoderm pro-



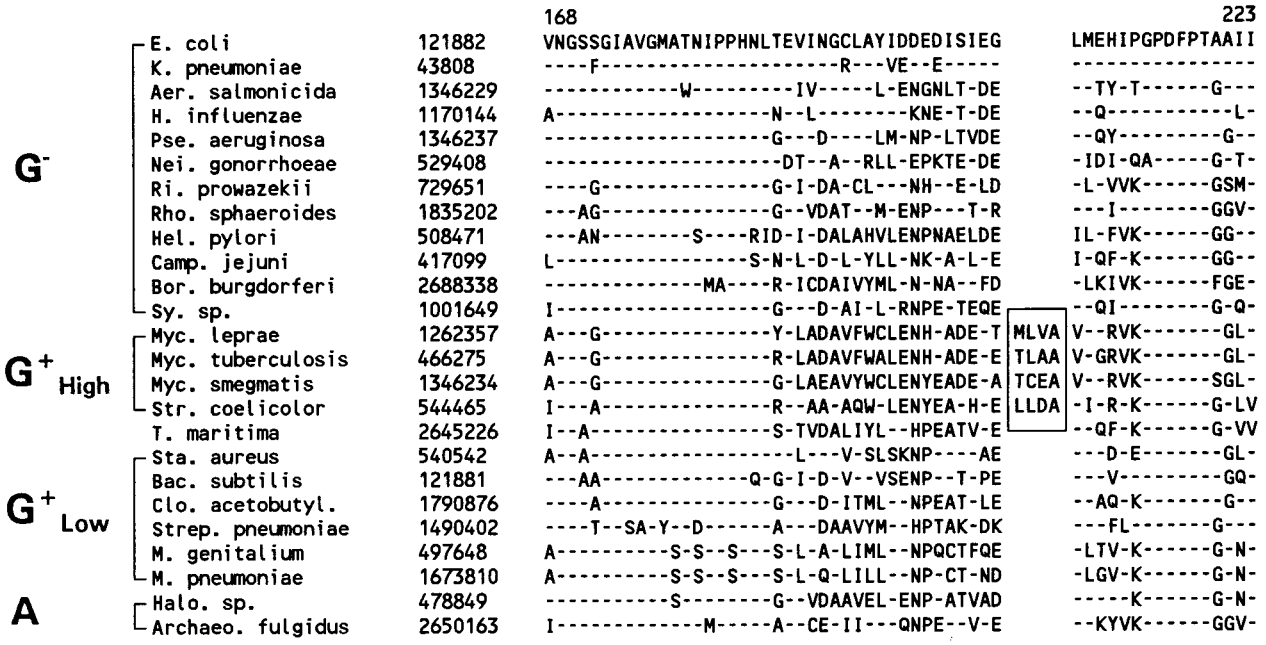
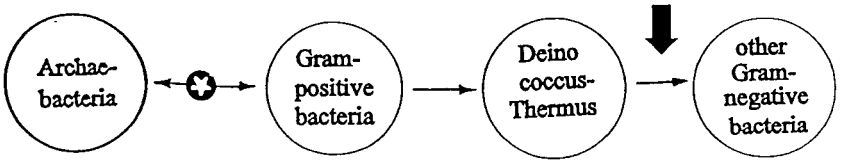


FIG. 15. Signature sequence (boxed) in the DNA gyrase A subunit which is specific for the high-G+C gram-positive group. As indicated in the top diagram, this signature was probably introduced in the branch leading to this group.

karyotes (i.e., archaebacteria and gram-positive bacteria) to the very early stages in the development of gram-negative (diderm) bacteria. I now present evidence that among the gram-negative bacteria, cyanobacteria constitute one of the deepest-branching divisions specifically related to the *Deinococcus-Thermus* group. The signature sequences that are helpful in establishing the phylogenetic position of cyanobacteria are present in a number of different proteins. These proteins include FtszA, in which a 1-aa insert is present in various proteobacteria and spirochetes but not in any of the archaebacteria, gram-positive bacteria, *Deinococcus*, cyanobacteria, and chloroplast homologs (Fig. 17a). Likewise, in the glutamate dehydrogenase (GDH) sequences, a 3-aa insert is present in various proteobacteria, bacteroides, and spirochetes but not in any cyanobacteria, *Deinococcus*, gram-positive bacteria, or archaebacteria (Fig. 17b). The species distribution of these signatures indicates that they were introduced after the evolution of cyanobacteria in a common ancestor of the various other divisions of gram-negative bacteria.

The signature sequences in the above proteins, which define a clade consisting of archaebacteria, gram-positive bacteria, *Deinococcus-Thermus*, and cyanobacteria, reinforce the view that archaebacteria and gram-positive bacteria are close relatives and that within gram-negative bacteria, cyanobacteria constitute one of the deepest-branching lineages. This inference is in accordance with the results of detailed phylogenetic studies based on a number of different gene sequences including Hsp70 (99, 103), Hsp60 (96, 98, 246), RecA (55, 131, 247), and 16S (33, 250) and 23S (48) rRNAs. In the case of Hsp70 sequences, for which sequence information is available from most bacterial phyla, a strong affinity of cyanobacteria to the *Deinococcus-Thermus* group was observed in both neighbor-joining and parsimony trees (bootstrap values, >99%) (103). Furthermore, a clade consisting of *Deinococcus-Thermus* species and cyanobacteria showed the deepest branching within gram-negative eubacteria and exhibited the closest relationship to the gram-positive bacteria. Similarly, a strong affinity of gram-positive bacteria to cyanobacteria (grouping in 99% of bootstraps) has been observed in the phylogenetic trees based on the GroEL-Hsp60 family of protein sequences (Fig. 11) (96, 98, 246). The members of other divisions (such as green nonsulfur bacteria and spirochetes) branched after this clade. A close relationship of cyanobacteria to the gram-positive bacteria has also been proposed based on the significant segment pair alignment scores in the RecA protein sequences (131) and



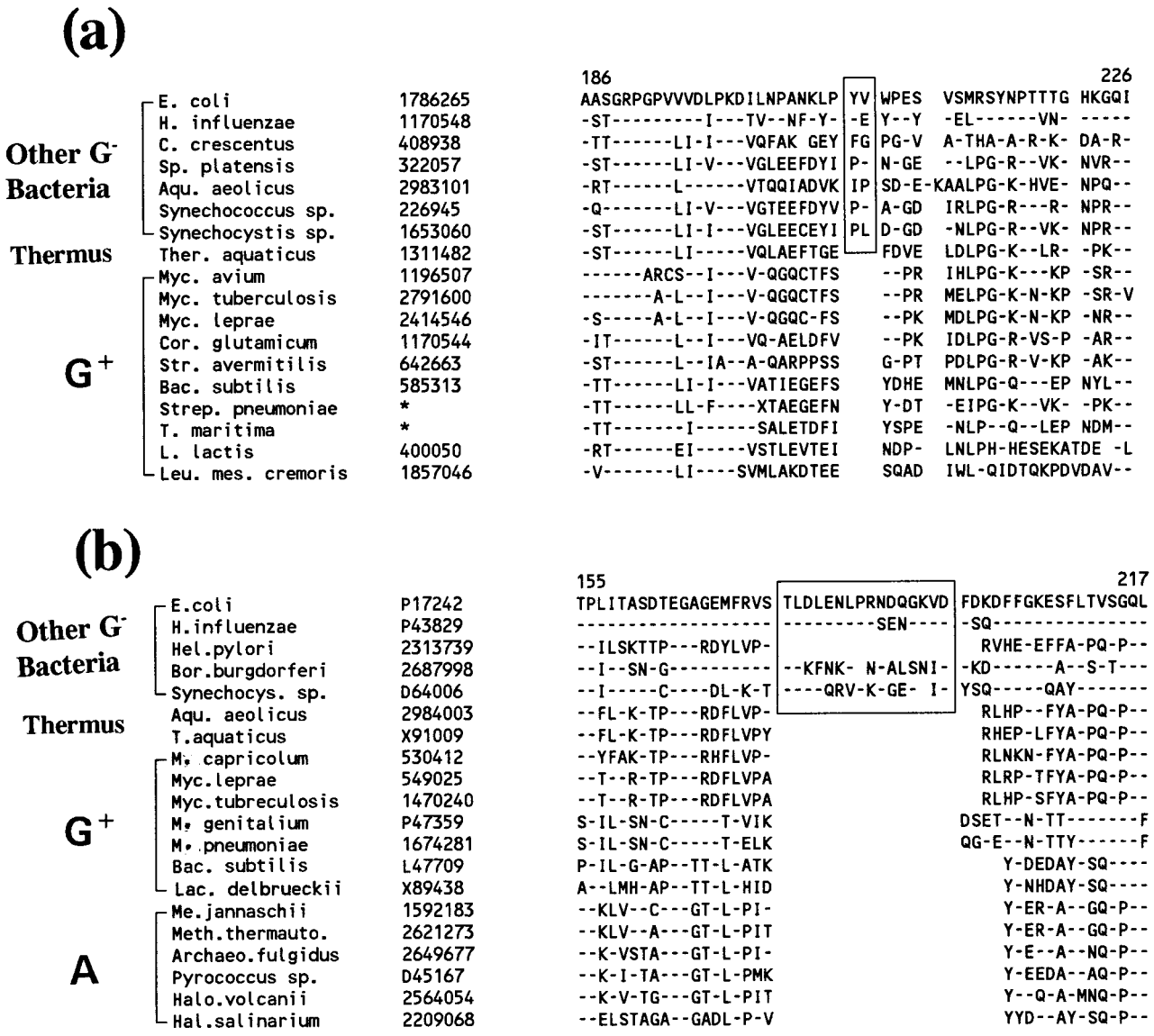


FIG. 16. Signature sequences (boxed) in acetolactate synthase (a) and asparginyl-tRNA synthetase (b) showing a grouping of the *Deinococcus-Thermus* species with archaebacteria and gram-positive bacteria. Similar to the Hsp60 protein (Fig. 10), these signatures were introduced in an ancestral gram-negative lineage after the branching of the *Deinococcus-Thermus* group.

the presence of a G residue at position 1207 in these groups in the 16S rRNA (250).

In addition to these sequence signatures, which are useful in understanding the evolutionary relationships of *Deinococcus-Thermus* and cyanobacteria to other prokaryotes, a number of other proteins contain yet another kind of signature sequence that is unique to only the *Deinococcus-Thermus* group and cyanobacteria and is not found in any other prokaryotes. In the DnaJ-Hsp40 family of proteins, the homologs from *Deinococcus-Thermus* and cyanobacteria contain a large deletion (68 aa) which removes the four cysteine-rich repeat domains that are present in all other prokaryotic and eukaryotic homologs (Fig. 18a) (27). Likewise, in the elongation factor EF-Ts sequences, the homologs from *Deinococcus-Thermus* and cyanobacteria harbored a deletion of 55 aa that is not present in other prokaryotes (Fig. 18b). Two other proteins where signatures unique to only these groups of prokaryotes are found are the protein synthesis elongation factor EF-Tu (Fig. 18c) and DNA polymerase I (Fig. 18d) (106). The presence of these uniquely shared sequence signatures in the *Deinococcus-Thermus* and

cyanobacterial phyla provides evidence of a close and specific evolutionary relationship between these two groups. These results also suggest that these two groups of organisms had a common ancestor exclusive of all other prokaryotes. However, this inference is difficult to reconcile with the signature sequences in other proteins (Hsp60, acetolactate synthase, and asparginyl-tRNA synthetase [Fig. 10 and 16]), which indicate that cyanobacteria and other gram-negative bacteria had a common ancestor exclusive of the *Deinococcus-Thermus* group and that the *Deinococcus-Thermus* lineage is more ancestral than cyanobacteria. To account for these observations, it is necessary to postulate that cyanobacteria and the *Deinococcus-Thermus* group are themselves not the direct ancestor of other gram-negative bacteria but branched off from the early ancestors as shown in the diagram in Fig. 18. Furthermore, to explain the presence of common sequence signatures in these groups that are not found in any other prokaryotes, it is necessary to postulate that some lateral gene transfers have occurred between these groups, as shown by the thin dashed arrow in Fig. 18. The possible significance of such lateral gene transfer events is discussed below (see "Evolutionary relationships within prokaryotes: an integrated view based on molecular and phenotypic characteristics").

Signature Sequences Defining Proteobacteria and Some of Their Subdivisions

The proteobacteria, named after the Greek god Proteus and meaning "capable of assuming many different shapes" (225), comprise one of the largest divisions among gram-negative bacteria. This group of bacteria, also called the "purple bacteria and relatives," exhibits diverse properties and is currently defined based mainly on the 16S rRNA phylogeny and signature sequences. Proteobacteria comprise more than 200 genera and have been divided into at least five subclasses: alpha through epsilon (177). However, the taxonomic relationship among proteobacteria, which include a very complex assemblage of phenotypic and physiological attributes, remains ill defined and has been a cause of concern (177).

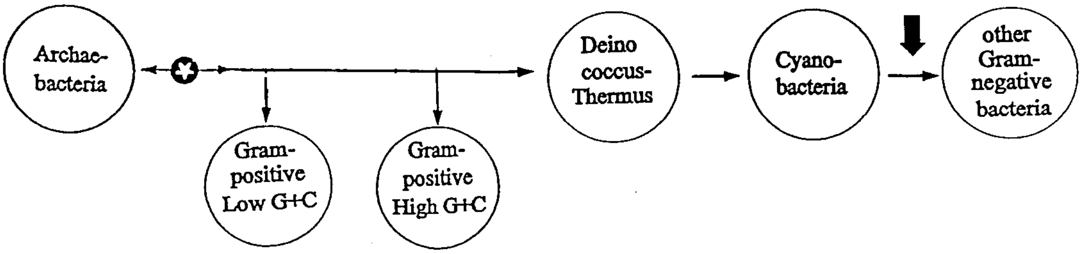
In the present work, although the signature sequences that may be useful in defining proteobacteria and its subclasses have not been examined in detail, I have identified some signature sequences that are useful in defining the proteobacterial group and some subdivisions within it. The first signature that is present in all proteobacteria examined, including members of all five subdivisions, consists of a 2-aa insert in the Hsp70 sequences (Fig. 19a). Interestingly, this signature is also present in the Hsp70 homolog from *Thermomicrobium roseum*, which is a member of the division "green nonsulfur bacteria and relatives." The branching position of the green nonsulfur bacteria, which consists of only a few species (*Thermomicrobium*, *Herpetosiphon*, and *Chloroflexus* species) in different phylogenies, including rRNA and Hsp70, has not been satisfactorily resolved (103, 184, 250). Hence, the above signature sequence, which is uniquely shared by various proteobacteria and a green nonsulfur bacterium but not by any other eubacterial groups, including cytophagas, flavobacteria, chlamydiae, spirochetes, cyanobacteria, *Deinococcus*, *Thermus*, and *Aquifex*, provides the first reliable evidence that some members of the green nonsulfur group of bacteria show a specific relationship to proteobacteria compared with the other divisions of eubacteria. The second signature sequence consists of a 4-aa insert in alanyl-tRNA synthetase, which is present in various species examined from the alpha, beta, gamma, and epsilon subdivisions of proteobacteria (Fig. 19b). Thus far, no sequences are available for this protein from members of the delta subdivision or green nonsulfur bacteria. These signature sequences were probably introduced into a common ancestor of proteobacteria after the branching of chlamydiae, cytophagas, and related species (Fig. 19a), and they could be used to define the proteobacterial group.

Another group of signature sequences that I have identified provides evidence that the members of the beta and gamma subdivisions are distinct from those of the other three subdivisions. The first of these signature consists of a 4-aa insert in a highly conserved region of Hsp70 (Fig. 20a) that is found uniquely in all members of the beta and gamma subdivisions that have been examined. A 1-aa insert in a highly conserved region, which is specific to only the members of the beta and gamma subdivisions, is also present in DNA gyrase A (Fig. 20b). A close affinity of the bacterial species for the beta and gamma subdivisions has also been observed in phylogenetic trees based on a number of genes and proteins, including SSU and LSU rRNA (48, 184, 250), Hsp60 (96, 246), Hsp70 (56, 103), RecA (55, 131), and sigma 70 (94). The signature sequences described above could be used to define the proteobacterial group and to distinguish members of the alpha, delta, and epsilon subdivisions (as well as *Thermomicrobium roseum*) from those of the beta and gamma subgroups (Fig. 19 and 20). These two proteobacterial groups are referred to as proteobacteria-1 and proteobacteria-2, respectively, in the remainder of this review.

Nature of the Archaebacterial Group and Its Relationship to Gram-Positive Bacteria

One of the main premises of current evolutionary thinking is that archaebacteria comprise a monophyletic group and that they are completely distinct from other prokaryotes (54, 183, 187, 250, 258). The uniqueness of archaebacteria is indeed supported by signature sequences in a number of different genes and proteins and by the major differences seen between archaebacteria and eubacteria in the information transfer processes (i.e., replication, transcription, and translation) (9, 47, 54, 144, 158, 183, 197). Several other characteristics of archaebacteria, including the unusual ether-linked nature of their membrane lipids (127, 133), are also consistent with this view. While these molecular features and characteristics point to the differences between archaebacteria and other prokaryotes, it is essential that we critically examine this relationship to understand the significance of these differences and the overall relationship of archaebacteria to other prokaryotes.

Within archaebacteria, phylogenetic analyses based on rRNA and EF-1 and EF-2 sequences have identified two main groups: *Crenarchaeota* and *Euryarchaeota* (149, 153, 184, 198, 250, 258). These groups are also distinguished from each other based on a signature sequence in the EF-1α/Tu protein, identified by Rivera and Lake (198). The *Crenarchaeota* consist almost exclusively of sulfur-dependent thermoacidophilic archaebacteria (250, 251, 258), and these genera are referred to as "eocytes" by Lake and coworkers (149, 154, 155, 198). The *Euryarchaeota* are phenotypically very diverse and have no specific physiological attribute. As indicated by Woese (253), this group is a "potpourri of all the archael types" and contains members from all phenotypically diverse archaebacteria including extreme halophiles, methanogens, sulfate-reducing archaebacteria, and sulfur-dependent thermoacidophilic archaebacteria. It is important to note that within the *Euryarchaeota*, members of different physiologically diverse archaebacterial phenotypes (halophiles, methanogens, thermoacidophiles, etc.) are not resolved from each other but show polyphyletic branching within three main clusters of methanogens (the *Methano-*



(a)

```
                                                  221                                                      274
Other G- Bacteria
 E.coli                 P06138     GHAMMGSGVAS GEDRAEEAAEMAISSPLLE D IDLSGARGVLVNITAGFD LRLDE
 H.influenzae           P45069     -Q--I-F-S-V --G------RL-VRND--- I K---N-Q-I-------M- -VFE-
 Azo.vinelandii         1518099    -M----T-F-- -PN--R--T-A--RN---- - VH-Q----I-------P- -S-G-
 Buch. aphidicola       2738589    -Y----T-IS- --N-------I-------- - ------------------ -K---
 Pse.putida             U29400     -M----T-C-- RPN--R--T-A--RN---- - VN-Q----I-------P- -S-G-
 Pse.aeruginosa         P47204     -M----T-C-- -PN--R--T-A--RN---- - VN-Q----I-------P- -S-G-
 Nei. gonorrhoeae       2494600    -I------Y-Q -I---RM-TDQ-------D - VT-D----------TAPGC-KMS-
 Nei.meningitidis       1657694    -I------Y-Q -I---RM-TDQ-------D - VT-D----------TAPGC-KMS-
 Wolbachia sp.          1169772    -K--I-T-E-E -----IS---A---N---D N VSMK--Q-I-I---G-G- MT-F-
 Bar. bacilliformis     2253395    -R----T-E-- --G--LA---A--AN---D E TSMC----L-IS--G-R- MT-F-
 A. tumefaciens         2465465    ARP---T-E-- -PA--MQ---A--AN---D E TSMK--Q-L-IS--G-R- -T-F-
 R.meliloti             P30327     -R----T-E-T --N--ML---A--AN---D E VSMR--K----S-SG-M- MT-F-
 C.crescentus           U40273     -K----T-EGT A----LM--QN--AN---D E VS-K--KA----V-G-M- MT-L-
 Hel. pylori            2494599    -F-L--I-E-T --ES-KL-VQN--Q----D - ASIE--KSII-FFEHHP- YPMMA
 Aqu. aeolicus          2983170    -LSII-M-EGR -DEK-DI-V-K-VT----- G NT-E---RL--T-WTSE- IPY-I
 Tre. pallidum          *          -Y-LI-V-EGE --N--VD--TA--NN---- E TRIE--TRL--AVRGSEN -SMG-
 Bor.burgdorferi        U43739     -D-L--I-YGK --N--VDRRTS---N---- E VRIE-SK-L---V-G-D- FS-L-
Deinococcus Cyanobacteria
 Sy.sp.PCC6803          2149909    -S-L--I--G- -KS--K---TA-------- SSIQ--K--VF-V-G-T- -T-H-
 Anabaena sp.           P45482     -S-L--I--S- -KS--R---IA-------- CSIE-----VF---G-S- -T-H-
 Ara.thaliana (c)       U39877     -T--L-V--S- SKN-------Q-TLA--IG SSIQS-T--VY---G-K- IT-Q-
 D. radiodurans         *          -TVL--I-AGR -DKM-----MS--H----- RGIE---RI---VTG-Y- -SMTD
G+
 Sta.aureus             P45498     -S-L--I--S- --N--V---KK-------- TSIV--Q---M---G-ES -S-F-
 Bac.subtilis           P17865     -S-L--I-I-T --N--A---KK-------- AAID--Q---M---G-TN -S-Y-
 Ent.faecalis           2149909    -T-L--I---- --E-VI--TKK-------- TSID--EQ--L---G-L- MT-F-
 Ent. hirae             2665345    -T-L--I---- --E-VI--TKK-------- TSID--EQ--L---G-L- MT-F-
 M.pulmonis             U34931     -D-LI-I-R-- -K---VK--IH-----II- TSIQ--SHTII---GSAN -T-T-
 Cor.glutamicum         1769961    -S-L--V-S-R -DN-VVS-T-Q--N----- ATMD--T---LSFAG-S- -G-M-
 Myc.tuberculosis       2104328    -T-L--I-S-R --G-SLK---I--N----- ASME--Q---MS-AG-S- -G-F-
 Str.coelicolor         P45500     -S-L--I-S-R -D---VA------------ ASID------LS-SG-S- -G-F-
 Str.griseus            P45501     -S-L--I-S-R -D---VA------------ ASID------LS-SG-S- -G-F-
 T.maritima             2104497    -A-IL-I--GK --H--R---KK-ME-K-I- HPVEN-SSIVF----PSN I-ME-
A
 Hal.volcanii           U37584     -V--I-L-ESD S-SK-Q-SVKS-LR----D V-I---NSA---V-G-S- MSIE-
 Halo. volcanii         2494606    -V--I-L-ESD S-SK-Q-SVKS-LR----D V-I---NSA---V-G-S- MSIE-
 Hal.salinarium         U32860     -V-V-LV-ETQ DKNKTN-VVKD-MNH---D V-YR--S-G--H--G-P- -T-K-
 Th. acidophilum        2724096    -V-LI-M-QSKK-G--IMT-L-E-LKPR-ID V-V-T-KDCVFK-I-PP- ITVS-
 Archaeo. fulgidus      2650085    -V--I-L-E-- ---K-A-SVRK-LK----D V-V---KAA---V-G-P- MTIE-
 Py.woesei              U56247     -V--I-I-ESD S-K--L----Q-LN----D V-I---S-A-IH-SGA - VK-E-
 Met.thermoauto.        2622805    -M--I-M-E-E SG---L-SVYE-LN----D L-I-N---A-I--SGSS- -T-Q-
 Me. jannaschii         2144960    -L--I-I-ESD S-K--K--VS--LN----D V-ID--T-A-IHVMGPE- -T-E-
```

(b)

```
                                306                                                                    369
A
 Hal.salinarium       49046     VDVLIPAALGNVITKENAEAIAA       DLVVEGANGPTTSTADSILADRDVAVIPDILANAGGVTVSY
 Sul.solfataricus     243120    C-I------E---N-F--PKVK-       K-I-------L-AD--E-MRQ-GI--V----------VG--
 Sul.shibatae         403324    C-I-----VE---N-F--PKVK-       K-I-------LAAD--E-IKQ-GIV-----------VG--
 Thermo.litoralis     310891    ----A-S-IEE----K--DN-K-       KI-A-L------PE--E--YEKGILI---F-C---------
 Py.furiosus          1122753   ----A---IEE----K--DN-K-       KI-A-V----V-PE--E--FEKGILQ---F-C---------
 Py. horikoshii       2828004   ----A---IEE----K--DN-K-       KI---V----V-PE--E--FEKGILQ---FLC---------
 Py.endeavori         464224    ----A---IEE----K--DN-K-       KI-A-V----V-PE--E--FEKGILQ---F-C---------
 Archaebac. ES4       A47410    ----A---IEE----K--DN-K-       KI-A-V----V-PE--E--FEKGILQ---F-C---------
G+
 Clo.difficile        144820    --IV-----E-S----V--S-K-       K--C-A------PE--EVF-E-GIVLT----T---------
 Bac.subtlis          413999    C-I-V---IS-Q--AK--HN-Q-       SI---R------ID-TK--NE-G-LLV-----S--------
 T.maritima           1743418   --I-V----EGA-HAG---R-K-       KA----------PE--E--SR-GIL-V--------------
 Pep.asaccharo.       150670    Y-IIV----E----G-R-KT-N-       K--C-A------PEG-KV-TE-GINLT----T-S---L---
Deinococcus Cyanobacteria
 D. radiodurans       *         C--------EKQ--LQ--DK-R-       R----------IPA--DL--QKG-T-V--V-------S---
 Sy.sp. PCC6803       1006751   ---------E-Q--RD--DQVR-       RYIF-V------TA--D---SKGIY-F----V---------
Other G- Bacteria
 Po.gingivalis        150842    --FAM-C-TQ-EMNL-D-KTLHK  NGV  T--A-TS-MGC-AE-SEYYVANKMLFA-GKAV-----SC-G
 E.coli               146124    --IAL-C-TQ-ELDVDA-HQLI-  NGV  KA-A----M---IE-TELFQQAG-LFA-GKA------AT-G
 Sal.typhimurium      154085    --IAL-C-TQ-ELDVDA-RVLI-  NGV  KA-A----M---IE-TDLFLEAG-LFA-GKA------AT-G
 H.influenzaea        1222106   --IAL-C-TQ-ELELSD-QRLIK  NGV  K--A----M---IE-TEA-LAA--LFG-GKA------AT-G
 Hel. pylori          2494098   CFAAF-S-TE-ELSVLD-KTLLS  NGC  KC-A----M-SSNE-IGLFLQAKISYGIGKA------S--G
 Bact.thetaiota.      1772847   C-IAL-S-TQ-ELNGDH-RQLV-  NGC  IA-S----M-S-PE-VRVFQ-AKILYA-GKA------S--G
 Bact.fragilis        1685286   A-IAL-C-TQ-ELNG-D-KNLID  NNV  LC-G-IS-MGC-PE-IDLFIEHKTMYA-GKAV-----AT-G
 Prev.ruminicola      1772845   A-IAT-C-TQDE-NEAE-KTLI-  NGV  FA-S----M--EPA-IKVFQ-AKILYC-GKAS-----AT-G
```

coccales, the *Methanobacteriales*, and the *Methanomicrobiales*) (184, 250–253). Likewise, the *Crenarchaeota* also does not unite all thermoacidophiles, and members of the orders *Thermoplasmales*, *Thermococcales*, and *Pyrodictales*, which are sulfur-dependent thermoacidophilic archaebacteria with similar phenotypes to *Crenarchaeota*, branch within the *Euryarchaeota* group (184, 250–253). Thus, the two main archaebacterial groups are merely phylogenetic constructs and they do not separate or cluster the diverse groupings of archaebacteria.

Of the two archaebacterial groups, *Crenarchaeota* has been proposed to be ancestral (253, 258). This suggestion is based on the observation that both archaebacterial groups contain members which are thermophilic, anaerobic, and sulfur metabolizing, and hence these characteristics, which are common in most members of the *Crenarchaeota*, are ancestral (253, 258). Lake and coworkers (148, 150, 155) have reached a similar inference independently. Based on the observation that ribosomes from this group of archaebacteria had certain distinctive features that were not present in other groups of prokaryotes but were shared with the eukaryotic ribosomes, Lake has proposed that the traits of this group are primitive and calls them "eocytes" (meaning dawn+cell) (148, 150, 155). Lake's proposal divides the prokaryotes into two groups, one consisting of only eocytes and the other encompassing all of the halophiles, methanogens, and different divisions of eubacteria (148–150, 155). However, the view that *Crenarchaeota* is the ancestral lineage of the two archaebacterial groups is not supported by the signature sequence present in the EF-1α/Tu protein. Rivera and Lake (198) have described an 11-aa insert that is present in various members of the *Crenarchaeota* as well as in all eukaryotes but not in other prokaryotes. An alternate alignment of the same sequence region shown in Fig. 21 suggests that the length of the insert in *Crenarchaeota* may be only 7 aa rather than 11 aa as originally proposed (198). A vestigial insert of 2 to 3 aa is also present in the same position in some members of the *Euryarchaeota*. However, the length of the insert (7 or 11 aa) or the presence of a vestigial insert in some other archaebacterial species does not change the main inference to be derived from this sequence signature. The important point here is that based on evidence presented above, the root of the prokaryotic tree has been placed between archaebacteria and gram-positive bacteria. The fact that this insert is not present in any gram-positive bacteria or in members of the *Euryarchaeota* but is found only in the *Crenarchaeota* group of archaebacteria (Fig. 21) strongly indicates that the absence of this insert (common to all eubacteria and members of the *Euryarchaeota*) is the ancestral phenotype. Hence, of the two archaebacterial groups, *Euryarchaeota* is ancestral (Fig. 21 diagram).

Current evolutionary thinking based on SSU rRNA shows a monophyletic nature of the archaebacterial domain and led to the view that archaebacteria constitute the third domain or form of life, but this view is not universally supported by all gene and protein phylogenies. Alternate phylogenetic trees can be based on a number of different proteins including some of the most conserved proteins found in the biota (Hsp70, GS I, GDH, and the *hisC*, *hisF*, *hisH*, *trpB*, and *trpD* products); in these trees the various archaebacterial species do not form a monophyletic group but instead show polyphyletic branching within gram-positive bacteria (10, 21, 22, 99, 103, 104, 108, 240). In Hsp70 trees, homologs from halobacterial species branched with the high-G+C gram-positive group whereas homologs from some methanogenic (and often thermoacidophilic) archaebacteria grouped with the low-G+C gram-positive bacteria (108). The observed polyphyletic branching of archaebacteria with gram-positive bacteria has been shown to be reliable, and in studies where different alternative tree topologies were considered, a polyphyletic branching such as that observed was strongly preferred over a monophyletic grouping of all archaebacteria by different phylogenetic methods (73, 104, 108). Similar relationships are adduced from signature sequences in some proteins (Fig. 13). Additionally, in dihydroorotate dehydrogenase, various members of the low-G+C gram-positive bacteria as well as methanogenic and thermoacidophilic archaebacteria lacked a 2-aa insert that is present in a halophilic archaebacterium and mycobacterial species (Fig. 22). It should be noted that the G+C content of halophilic archaebacteria is in the range of 66 to 68% whereas that of methanogens is <50% and generally in the range of 30 to 40% (229). This 2-aa insert is also present in various gram-negative bacteria, supporting the evidence derived from other signature sequences (see "Signature sequences distinguishing between low-G+C and high-G+C gram-positive bacteria and pointing to a specific relationship of the latter group to the gram-negative bacteria") that within gram-positive bacteria, members of the high-G+C group are the closest relatives of the gram-negative bacteria.

The above observations raise important questions about the true relationship between archaebacteria and gram-positive bacteria. While some genes and proteins provide evidence that archaebacteria are distinct from other prokaryotes and the primary division within them is between *Euryarchaeota* and *Crenarchaeota*, for a number of other genes the archaebacteria do not form a monophyletic group but instead show polyphyletic branching within gram-positive bacteria, with the halophilic archaebacteria showing affinity for the high-G+C group and some methanogens branching with the low-G+C group. To explain these results, it is necessary to postulate that some lateral or horizontal gene transfers have taken place between these two groups of prokaryotes. Although the exact nature of the gene transfer events between these groups remains unclear, the observed results could be explained by two different scenarios (Fig. 23).

The first scenario (I in Fig. 23) assumes that archaebacteria are indeed a monophyletic group distinct from gram-positive bacteria. In this case, to explain the observed results, one has to postulate that genes for many of the proteins for which archaebacteria show a polyphyletic branching within gram-positive bacteria (e.g., Hsp70, GS I, GDH, the *hisC*, *hisF*, *hisH*, *trpB*, and *trpD* products, and dihydroorotate dehydrogenase) have been transferred from low-G+C gram-positive bacteria to methanogens and thermoacidophilic archaebacteria and from high-G+C gram-positive bacteria to the halophiles. At the same time, the corresponding genes from the archaebacteria (if any unique genes for these proteins were present in archaebacteria) have been lost. The alternate scenario (II in Fig. 23) to explain these results assumes that archaebacteria are indeed closely related to gram-positive bacteria, as suggested by some of the most highly conserved proteins, and that they may have evolved from specific members of low- and

FIG. 17. Sequence signatures in FtsZ (a) and glutamate dehydrogenase (b) showing the relatedness of cyanobacteria (and chloroplast homologs) to gram-positive bacteria and archaebacteria. As shown in the diagram above, these signatures (boxed) were probably introduced in a common ancestor of other gram-negative bacteria after the branching of cyanobacteria.

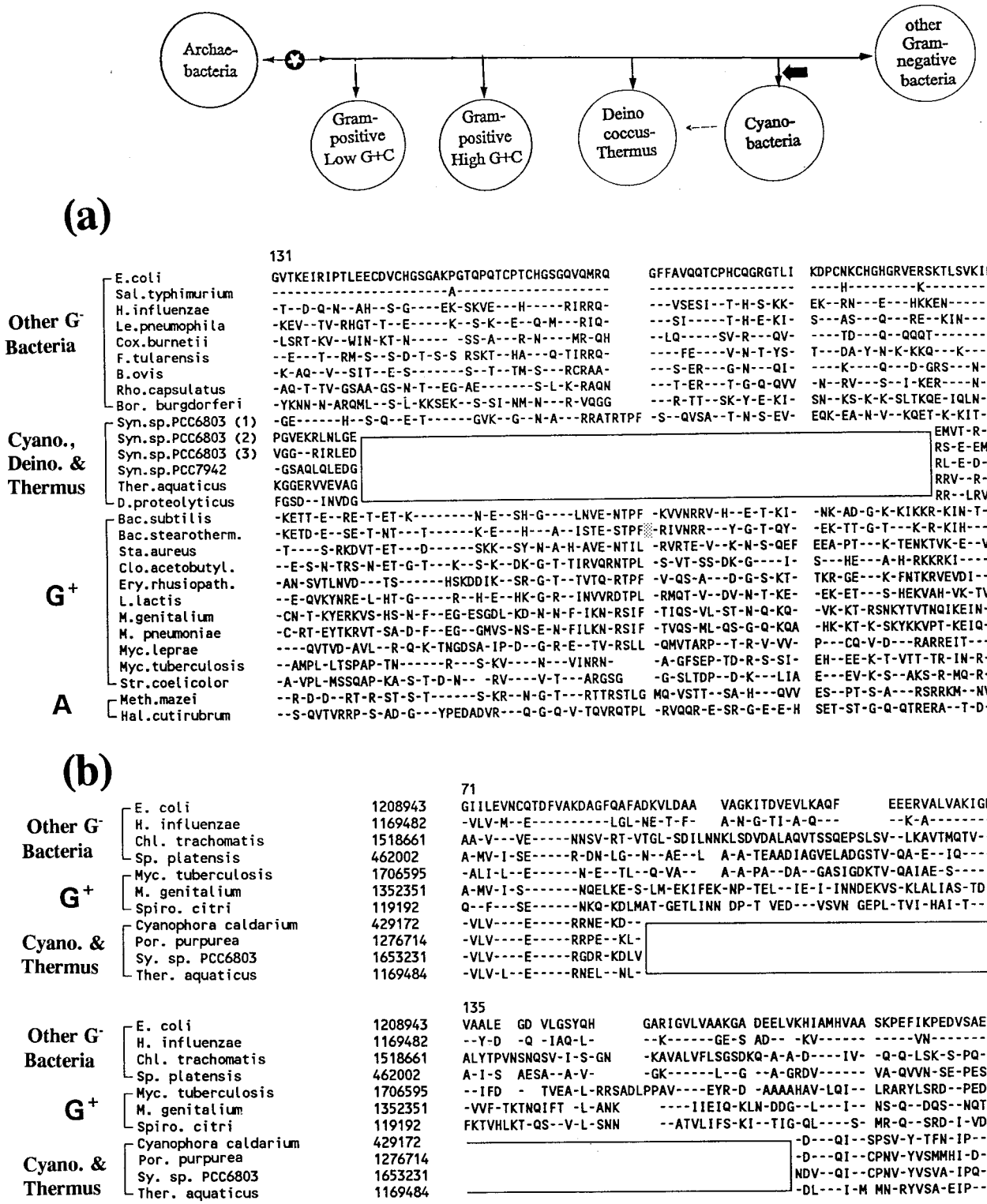


FIG. 18. Signature sequences in DnaJ (a), EF-Ts protein (b), EF-Tu protein (c), and DNA polymerase I (d) that are unique to only the *Deinococcus-Thermus* group and cyanobacteria. To explain the presence of these signatures (boxed), as well as those in Fig. 10 and 16, it was suggested that these signatures were introduced initially into the branch leading to cyanobacteria (thick arrow) and then laterally transferred to the *Deinococcus-Thermus* group (thin dashed arrow). The alternate possibility, that these signatures were first introduced into the branch leading to *Deinococcus-Thermus* and then transferred to cyanobacteria, is also possible. Panels b through d reproduced from reference 106 with permission of the publisher.

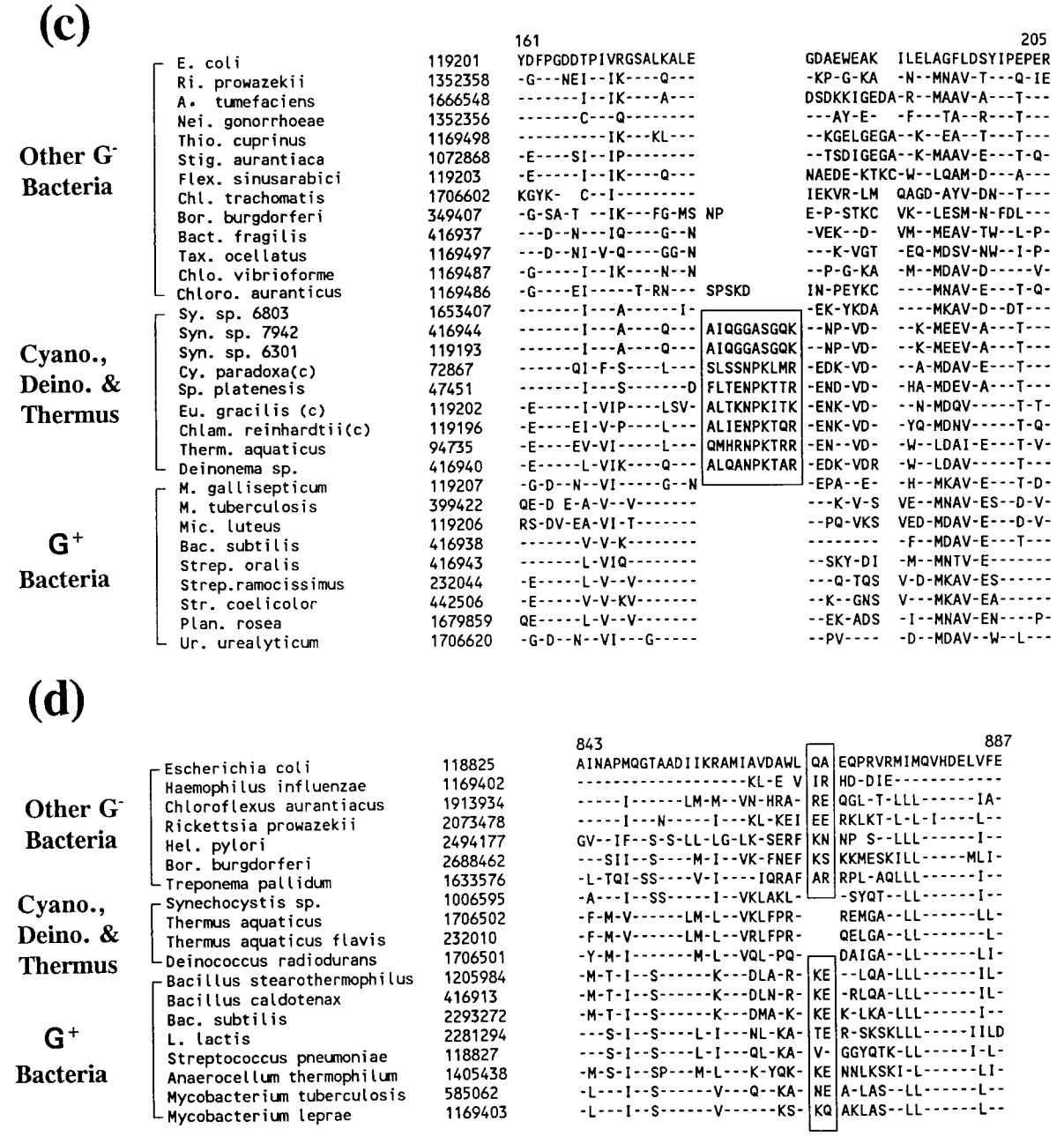


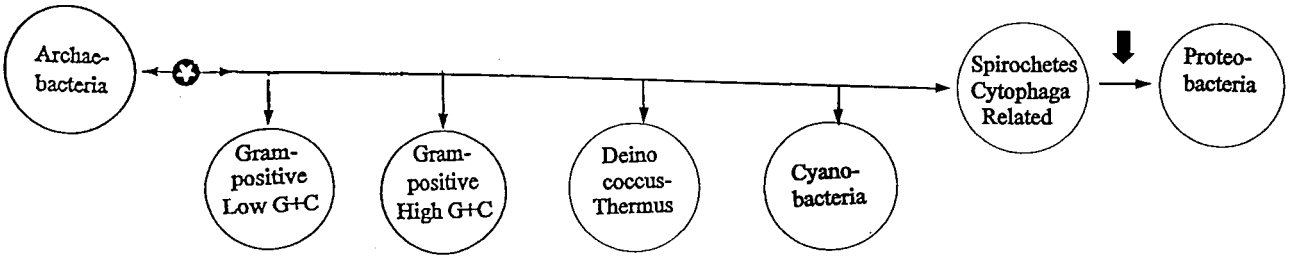
FIG. 18—*Continued.*

high-G+C gram-positive bacteria, as suggested by these phylogenies and signature sequences, as well as the G+C content of the halophilic (high-G+C) and methanogenic (low-G+C) archaebacteria (229). In this case, to account for the results from different gene phylogenies, one has to postulate that the genes for many functions that indicate a monophyletic nature of archaebacteria were transferred from one or more gram-positive bacteria that originally evolved such changes into others. This latter scenario, if true, suggests that the earliest prokaryote was a low-G+C gram-positive bacterium.

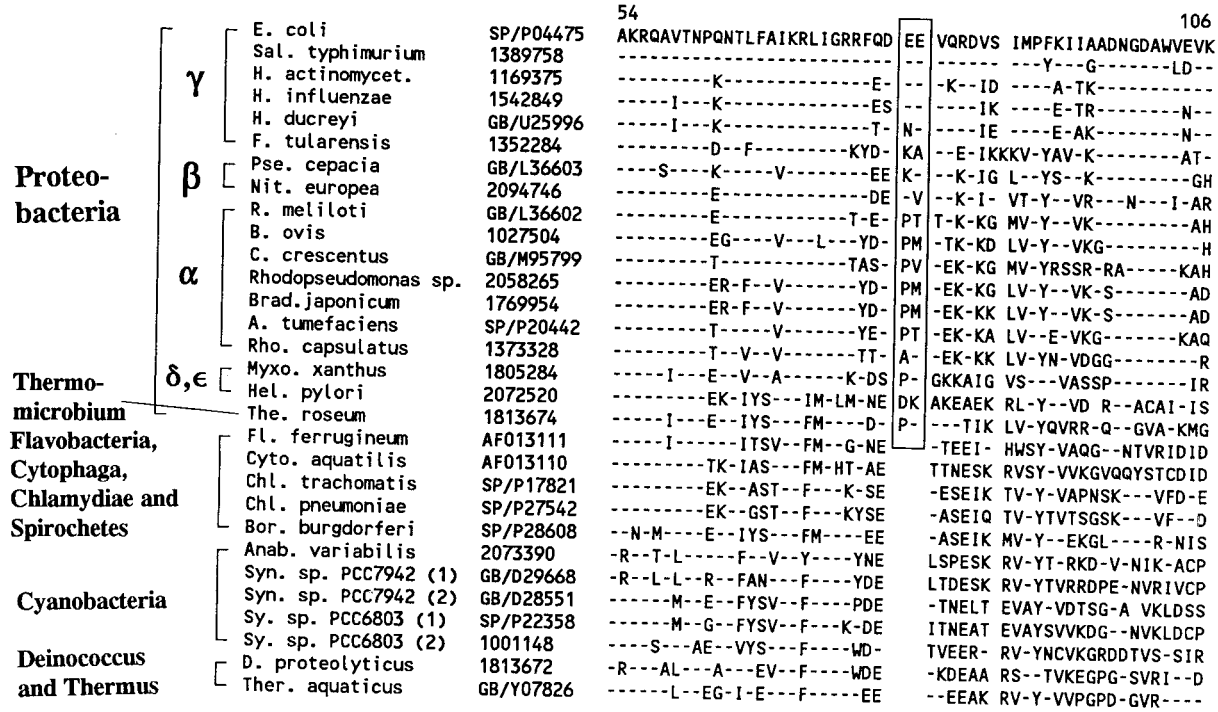
Both of these possibilities could explain the observed results, and neither should be dismissed a priori without serious consideration. In the past, supporters of the three-domain proposal have favored the first of these possibilities (7, 21, 53, 183, 216), and the alternate possibility has not been considered.

Possible Selective Forces Leading to Horizontal Gene Transfers

From the signature sequences that I have described thus far, it should be evident that the horizontal gene transfer between species is not all that common. If this was occurring commonly and indiscriminately between species, the clear distinction between different phylogenetic groups that we have observed based on signature sequences in different proteins would not have been possible. These results are at variance with the



(a)



(b)

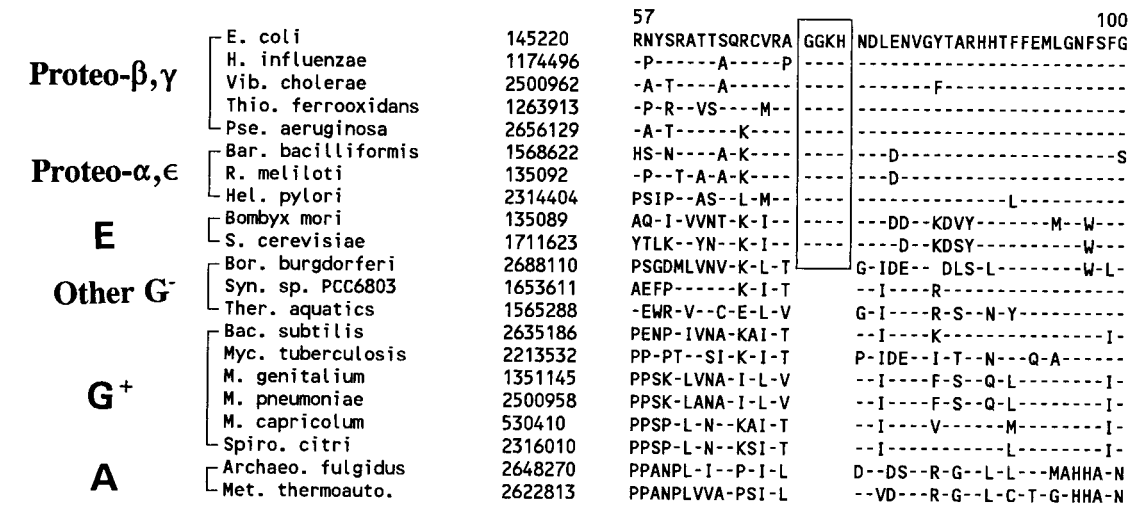
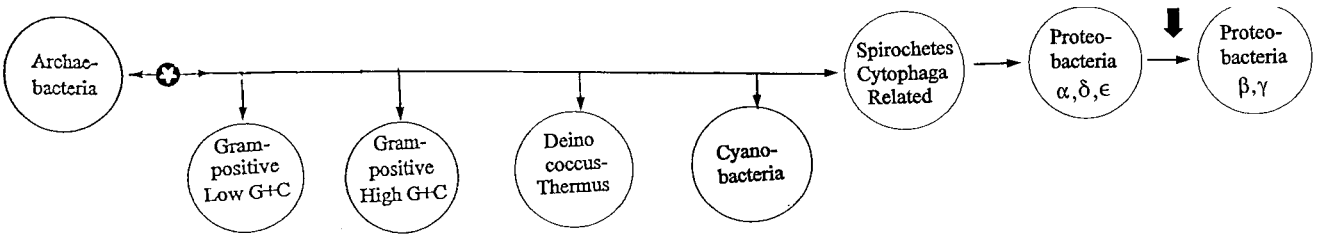
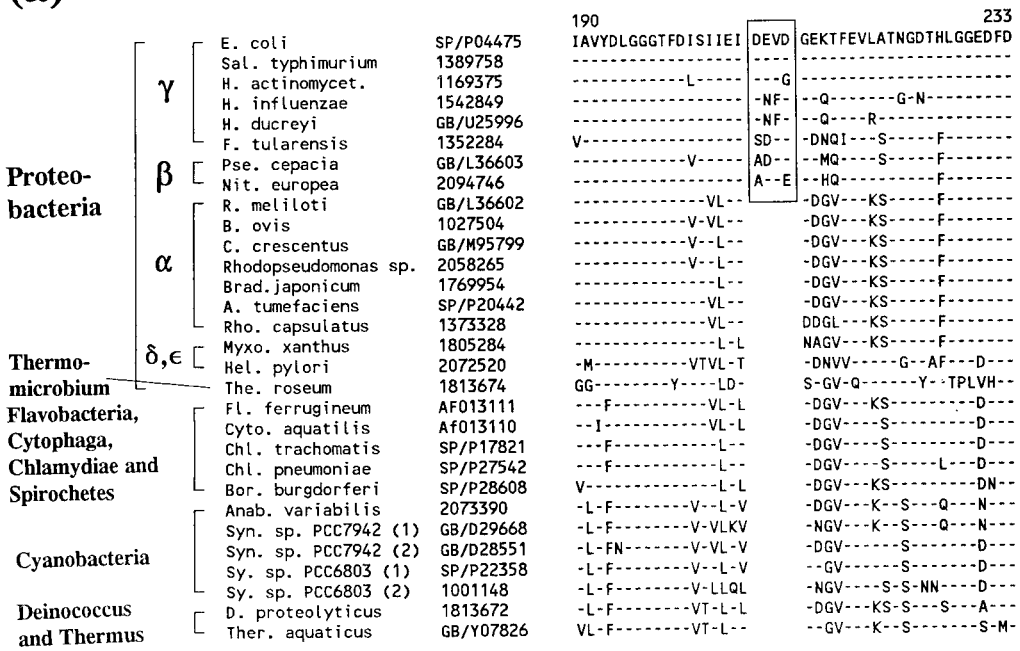


FIG. 19. Signature sequences (boxed) in Hsp70 (a) and alanyl-tRNA synthetase (b), defining and distinguishing proteobacterial group from all other divisions of prokaryotes.



(a)



(b)

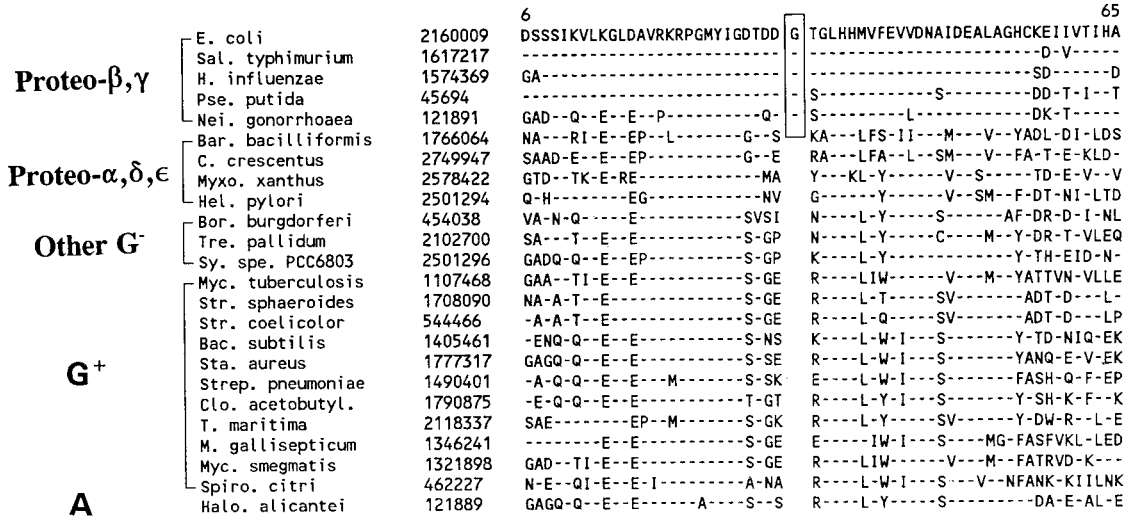
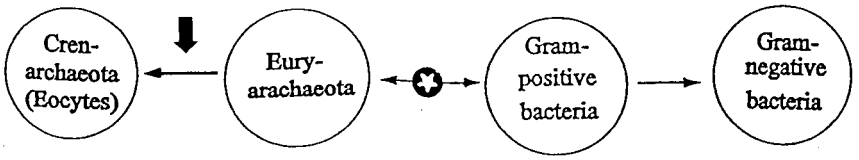


FIG. 20. Signature sequences in Hsp70 (a) and DNA gyrase B (b) which appear specific for the beta and gamma subdivisions of proteobacteria. These signature sequences (boxed), in combination with those in Fig. 19, could be used to define and distinguish between proteobacterial subdivisions alpha, delta, and epsilon (proteobacteria-1) and subdivisions beta and gamma (proteobacteria-2).



			98		131
Euryarchaeota	Thermo.celer	P17197	KNMITGASQADAAVLVVAVTDG		VMPQTKEHAFLA
	Hal.marismortui	P16018	----------N------AD--		-Q---Q--V---
	Hal. halobium	352354	----------N------AD--		-A---R--V--S
	Archaeo. fulgidus	2649659	---------------MD-VEK		-Q---O--I---
	Th.acidophilum	P19486	------T------I--ISAR--	EG	--E--R------
	Met. thermoauto.	2622158	------------------D--		--------V--S
	Me. jannaschii	2494244	----------------D-N-A	KTG	IQ---R--M---
	Me.vannielii	P07810	----------------N-D-A	KSG	IQ---R--V--I
	Py.woesei	P26751	------------------A---		------------
Crenarchaeota (Eocytes)	Sul.acidocaldarius	P17196	------------I---SAKK-	EYEAGMS	AEG--R--II-S
	Sul.solfataricus	P35021	------------I---SAKK-	EYEAGMS	-EG--R--II--
	Des.mobilis	P41203	------------I---SARK-	EFEAGMS	AEG--R---I--
Eukaryotes	En.histolytica	X83684	------T----V-I-I--AGT-	EFEAGIS	KNG--R--IL-S
	Rh.racemosus	P14865	------T----C-I-II-GGT-	EFEAGIS	KDG--R---L--
	S.cerevisiae	P02994	------T----C-I-II-GGV-	EFEAGIS	KDG--R---L--
	Human	P04720	------T----C---I--AGV-	EFEAGIS	KNG--R---L--
	Tomato	P17786	------T----C---IIDS-T-	GFEAGIS	KDG--R---L--
	Di.discoiedium	P18624	------T----C----I-SPT-	EFEAGIA	KNG--R---L--
Bacteria G^+ and G^-	Myc.leprae	P30768	-------A-M-G-I----A---		P----R--VL--
	Str.coelicolor	X77039	-------A-M-G-I----A---		P-------VL--
	Mi.luteus	P09953	-------A-M-G-I----A---		P-A--R--VL--
	T.maritima	M27479	-------A-M-G-I----A---		P----R--VL--
	Sp.platenesis	P13552	-------A-M-G-I---SAA--		P----R--IL--
	Bac.subtilis	P33166	-------A-M-G-I---SAA--		P----R--IL-S
	M.genitalium	P18906	-------A-M-GGI---SA---		P----R--IL--
	E.coli	P02990	-------A-M-G-I----A---		P----R--IL-G
	Pse.cepacia	P33167	-------A-M-G-I--CSAA--		P----R--IL--
	Bact.fragilis	P33165	---V---A-M-G-II---A---		P----R--IL--
	Syn.sp. PCC7942	416944	-------A-M-G-I---SAA--		P----R--IL--
	Chl.trachomatis	P26622	-------A-M-G-I---SA---		A-------IL--
	Bor.burgdorferi	P23125	-------A-M---I-L--ADS-		AE------LL--
	Ri.prowazekii	P48865	-------A-M-G-I---SAA--		P----R--IL--
	Chloro.auranticus	P42472	-------A-M-G-I---SAP--		P----R--IL--

FIG. 21. Excerpts from EF-1α/Tu protein sequences showing a conserved insert (originally identified by Rivera and Lake [198]) that is present in various *Crenarchaeota* archaebacteria (eocytes), as well as eukaryotic homologs, but absent in *Euryarchaeota* archaebacteria and eubacteria. This insert indicates that of the two archaebacterial groups, *Euryarchaeota*, are ancestral.

recently developing consensus that the horizontal gene transfer between species is very common (21, 118, 142, 143, 191, 236a, 260). For lateral or horizontal transfer of genes between species to occur generally two related conditions are required. First, the gene to be transferred should confer a selective advantage on the recipient species. Second, a strong selective environment favoring the growth and survival of the species containing the transferred gene should exist. To understand the nature of horizontal gene transfers that may have taken place in the past, it is necessary to consider or speculate about the selective forces that may have been operative or existed in the primitive environment. Of the two possible scenarios for gene transfer suggested above (Fig. 23), I cannot think of any strong selective advantage that transfer of genes such as the Hsp70, GS I, GDH, and dihydroorotate dehydrogenase genes, from gram-positive bacteria will confer on the recipient archaebacteria (i.e., scenario I). On the other hand, a number of observations can be cited which support the second scenario.

In this context, it is important to point out that the main differences between archaebacteria and gram-positive bacteria are with regard to the functions that are involved either in information transfer processes (9, 47, 54, 144, 183, 197) or in the synthesis of cell wall components and membrane lipids (127, 133). These processes provide the main targets for the action of many commonly used antibiotics, e.g., chloramphenicol, erythromycin, tetracycline, streptomycin, kanamycin, neomycin, rifampin, actinomycin D, mitomycin C, adriamycin, novobiocin, gentamicin, bacitracin, and polymyxins, produced by different genera of gram-positive bacteria (6, 14, 179). Table 2 gives the site of action and the source of producing organisms for several antibiotics. This list is not exhaustive, and there are hundreds of less well studied antibiotics, produced by different gram-positive bacteria, that act on these targets (79, 179, 204, 228, 229). The production of these antibiotics or secondary metabolites by the producing bacteria provides them with a great selective advantage over other biota. As noted by Cava-

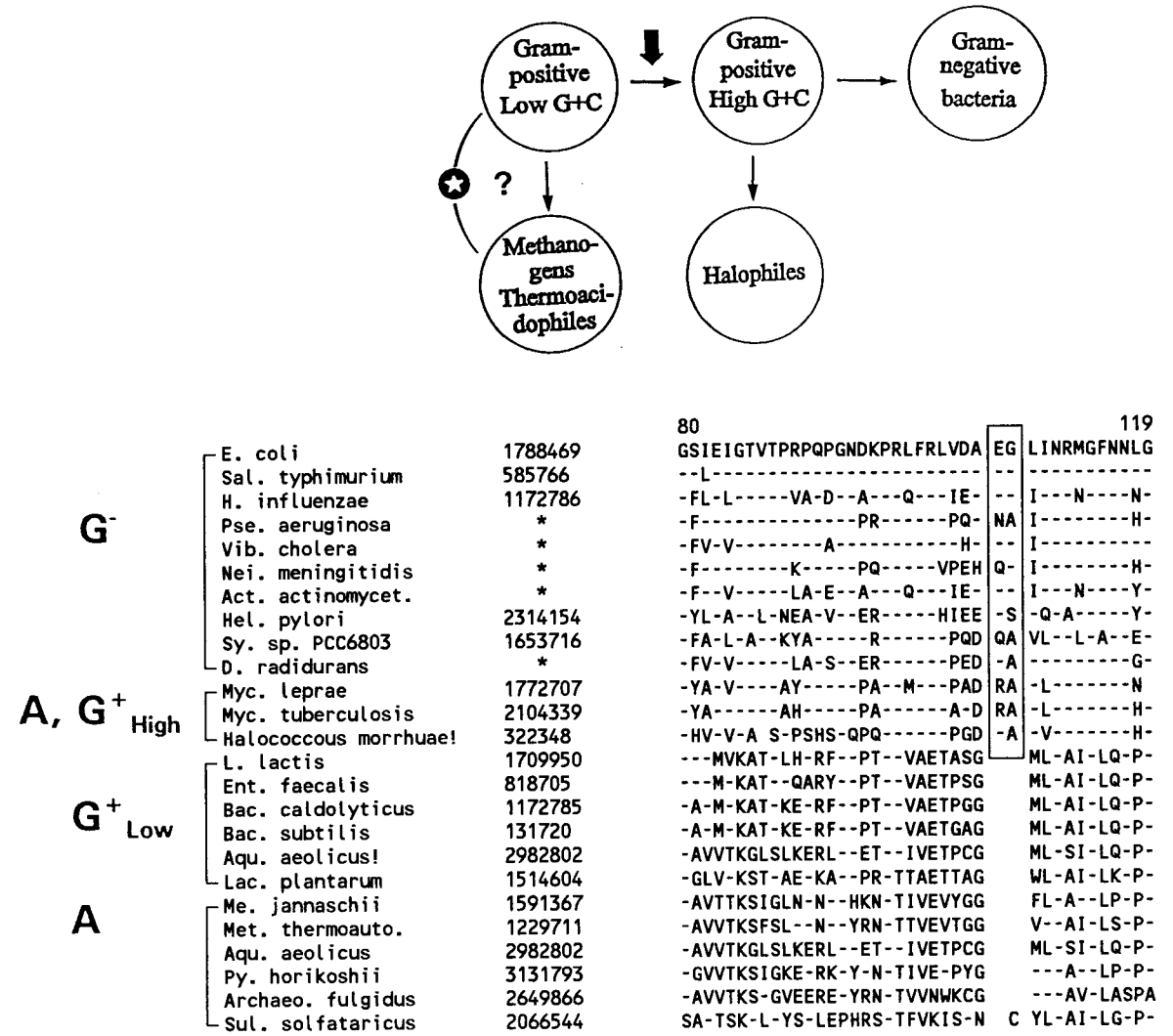


FIG. 22. Signature sequence in dihydroorotate dehydrogenase showing the relatedness of halophilic archaebacteria to the high-G+C gram-positive bacteria and of the methanogenic and thermoacidophilic archaebacteria to the low-G+C group. The thick arrow indicates that similar to other protein sequences (Fig. 13), this signature provides evidence that gram-negative bacteria are specifically related to the high-G+C gram-positive group. To explain the results with archaebacterial homologs, it is necessary to postulate either that there was a lateral gene transfer from high-G+C gram-positive bacteria to the halophilic archaebacteria or that the two groups of archaebacteria bear specific relationships to the two divisions of gram-positive bacteria (thin solid arrows). The question mark indicates that these results raise questions about the evolutionary relationship between archaebacteria and gram-positive bacteria.

lier-Smith (31): "Secondary metabolites (antibiotics) are most often beneficial to their producers as agents of the chemical warfare which is perpetually being waged against competitors, predators and parasites." Thus, it is quite likely that in the primitive environment some groups of gram-positive bacteria were producing antibiotics and others were sensitive to them, so that their survival was at stake. To survive in this environment, the sensitive bacteria had to undergo changes in the target sites of the above antibiotics so that their growth was no longer inhibited (6, 14, 44, 79, 179, 204, 223). There was thus very strong selection pressure on the genes that were the targets for these antibiotics to undergo changes to survive in the selective environment.

It is possible that under these conditions, after a long period of (repeated) selection in the primitive environment, assisted by every conceivable sort of stress (83), a resistant strain evolved that had undergone extensive changes in the genes that are the targets of the above antibiotics. This resistant strain may have been an ancestor of the present-day archaebacteria. Once a bacterium has developed a successful strategy to combat the effects of antibiotics, the other sensitive bacteria can readily acquire the resistance by means of genetic exchange or horizontal gene transfer from the resistant strain (36, 37, 44, 204, 223). As stated by Cohan (37): "Adaptations may be passed from one bacterial species to another, either by homologous recombination or by plasmid exchange. . . . The genes that can be transferred across taxa are necessarily a very small set that confer general adaptations which are not limited to the ecological and genetic context of a particular taxon (e.g., genes conferring resistance to widely used antibiotics). . . . A single genetic exchange in which an adaptation is transferred across taxa can change forever the course of adaptive evolution in the recipient species." This scenario can readily explain why in phylogenies based on such gene sequences as those encoding rRNA, EF-1α/Tu, EF-2/G, etc., phenotypically and physiologically diverse groups of monoderm prokaryotes (methano-

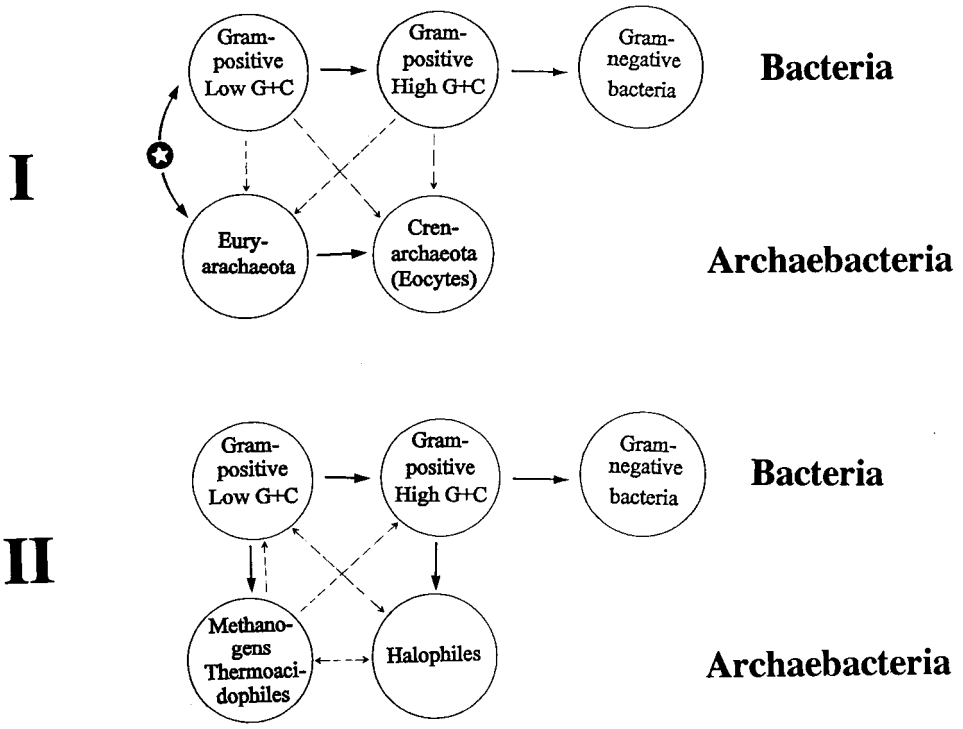


FIG. 23. Possible scenarios to explain the evolutionary relationship between archaebacteria and gram-positive bacteria. Scenario I assumes the archaebacteria to be monophyletic; to explain various other gene phylogenies where archaebacteria show polyphyletic branching within gram-positive bacteria (e.g., Hsp70, GS 1, GDH, dihydroorotate dehydrogenase), lateral transfer of genes from different groups of gram-positive bacteria to the archaebacteria (as indicated by thin dashed arrows) is postulated. Scenario II, on the other hand, suggests that the ancestral archaebacterial phenotype may have evolved from gram-positive bacteria (solid arrows) in response to antibiotic selection pressure and that the genes involved in antibiotic resistance (which may include many genes involved in the information transfer processes) were subsequently acquired laterally by other gram-positive bacteria to create additional monoderm prokaryotes with an archaebacterium-like genotype.

gens, halophiles, and thermoacidophiles) form a monophyletic group (i.e., archaebacteria) and also exhibit paraphyletic relationships within the *Euryarchaeota* division (184, 250–253). It has been noted by Woese (250) and others that the evolutionary distances in rRNA between some of the archaebacterial groups are very short: "Since their last common ancestor, archaebacterial rRNA on average have accumulated substantially fewer mutations than the rRNA of either eukaryotes or eubacteria" (186), leading to the inference that the archaebacterial lineage is slowly evolving (250, 258). However, the apparent slow evolution of the archaebacterial lineage could also be readily explained if many of the archaebacterial genes, rather than being ancestral, were acquired more recently by means of horizontal gene transfer between archaebacteria and Gram-positive bacteria. The exchange of genetic information in archaebacteria at a high frequency, at the high temperatures at which many of them grow, has been reported (93). Although the scenario presented here for the origin of archaebacteria is speculative, it is realistic, and it should be possible to test it experimentally.

The evolution of gram-negative bacteria, which possess an outer membrane, from gram-positive bacteria may also have been a defensive strategy on the part of some gram-positive bacteria to combat or at least minimize the effect of antibiotics (180a). As stated by Inouye: "The outer membrane serves as a selective barrier to the cell exterior. . . . gram-negative bacteria are more resistant to the actions of certain dyes, chemicals and antibiotics. This is because gram-negative bacteria have an outer membrane that prevents toxic compounds from entering the cells" (124). It is of interest in this regard that in the recently reported genomic sequence for the gram-positive bacteria *Bacillus subtilis*, a very large number of genes (77 in all) encoding the ABC transporter proteins were found (147). As noted in reference 147, these transporter proteins probably allow these bacteria to escape the toxic action of many compounds (antibiotics). Thus, the prokaryotic organisms have developed a number of different strategies to protect themselves from the toxic effects of antibiotics.

It should be clear from the above discussion that there is a lot to be learned about the relationship between the archaebacteria and gram-positive bacteria and the monophyletic and distinct nature of archaebacteria is far from established.

Evolutionary Relationships within Prokaryotes: an Integrated View Based on Molecular and Phenotypic Characteristics

Based on the phylogenies and signature sequences I have described thus far, the evolutionary relationship within the prokaryotic organisms that emerges is depicted in Fig. 24. The evolutionary relationship within the prokaryotic species is indicated to be a continuum, and the different groups shown in this figure appeared to have evolved from the common ancestor in the order shown. Of these groups, archaebacteria and gram-positive bacteria, the prokaryotes surrounded by a single membrane, are indicated to be the most ancient lineages within prokaryotes. The question whether the earliest prokaryote was a gram-positive bacterium, an archaebacterium, or a common ancestor from which both these lineages evolved independently is unclear at present. As discussed above, the answer to this question depends upon clarification of the evolutionary relationship between gram-positive bacteria and archaebacte-

TABLE 2. Sources and sites of action of some antibiotics[a]

Mechanism of action	Antibiotic	Producing organism
Inhibition of protein synthesis (acting on 30S ribosomal subunit)	Streptomycin	*Streptomyces griseus*
	Neomycin	*S. fradiae*
	Tetracyclines	*S. aureofaciens*
	Spectinomycin	*S. spectabilis*
	Gentamycin	*Micromonospora purpurea*
	Tobramycin	*S. tenebrarius*
	Pactamycin	*S. pactum*
	Kanamycin	*S. kanamyceticus*
Inhibition of protein synthesis (acting on 50S ribosomal subunit)	Erythromycin	*S. erytherus*
	Carbomycin	*S. halstidii*
	Chloramphenicol	*S. venezuelae*
	Lincomycin	*S. lincolenesis*
	Streptogramins	*S. graminofaciens*
Inhibition of synthesis or damage to cell wall	Penicillins	*Penicillium chrysogenum*
	Cephalosporins	*Cephalosporium acremonium*
	Bacitracin	*Bacillus subtilis*
	Vancomycin	*S. orientalis*
	Cycloserine	*S. garyphalus*
Inhibition of synthesis or damage to cytoplasmic membrane	Polymyxin	*B. polymyxa*
	Tyrothricin	*B. brevis*
	Gramicidin S	*B. brevis*
Inhibition of synthesis or metabolism of nucleic acids	Rifampin	*S. mediterranei*
	Novobiocin	*S. niveus*

[a] Compiled from information in references 6, 14, 44, 79, 179, 204, and 229.

ria. However, irrespective of whether archaebacteria constitute a monophyletic group distinct from other bacteria or whether they evolved from within the gram-positive bacteria, the inference that archaebacteria are more closely related to gram-positive bacteria than to gram-negative bacteria is supported by signature sequences in numerous proteins and by most of the gene and protein phylogenies. A specific relationship between archaebacteria and gram-positive bacteria is also strongly corroborated by the structural organization of their cells: within prokaryotes, only these two groups of organisms are bounded by a single lipid membrane (i.e., monoderm prokaryotes). Thus, the phylogenetic inferences based on macromolecular sequence data are in accord with the most important structural distinction seen within prokaryotes and there are no major conflicts between molecular phylogenies and phenotypic characteristics (Fig. 24), unlike previously (181, 252).

These results raise the important question of the primary division within the prokaryotes. The three-domain proposal divides prokaryotes into two primary groups: archaebacteria (*Archaea*) and eubacteria (*Bacteria*), and it does not recognize gram-negative bacteria (diderm bacteria with an inner and outer membrane defining a periplasm) as a distinct phylum. It is important to point out that the taxon *Archaea* has been defined only by biochemical and sequence characteristics and that its members show no unique morphological features by which they could be distinguished from gram-positive eubacteria (99, 258). Since the phylogenetic distinctness of *Archaea* is now highly questionable, and in view of the concerns raised that "It is not appropriate to separate kingdoms on any basis but a major, reasonably easily determined difference in organization" (175), I conclude that the basic premise of the three-domain proposal, i.e., that the primary division within prokaryotes is between *Archaea* and *Bacteria*, is not justified. In contrast to this proposal, the division of prokaryotes into two naturally defined, nonoverlapping primary taxa, *Monodermata* (prokaryotic cells surrounded by a single unit lipid membrane; includes all archaebacteria and gram-positive bacteria) and *Didermata* (prokaryotic cells containing both inner and outer unit lipid membranes enclosing a periplasmic compartment; includes all true gram-negative bacteria), is strongly supported by both morphological and molecular sequence characteristics (100, 101). Based on signature sequences, the monoderm prokaryotes could be divided into two main groups: gram-positive bacteria and archaebacteria. Any lateral gene transfer between these two groups of monoderm prokaryotes, as seems to have taken place, should not affect or influence their placement in the same taxon. Archaebacteria can be further divided into two subtaxa: *Euryarchaeota* and *Crenarchaeota* (eocyte), based on signature sequences in EF-1α/Tu (Fig. 21) (198). The signature sequences also support the division of gram-positive bacteria into two distinct group corresponding to low-G+C and high-G+C species (101). A clear phylogenetic distinction between the latter groups of species in the past has not been made. Further studies should clarify whether gram-positive bacteria contain additional groups which may include species such as *Thermotoga maritima*. It should be emphasized that although I have used the common names to designate various higher taxa and subtaxa within prokaryotes, these names do not constitute the defining characteristics of these groups. All of the taxa described here are defined based on specific signature sequences in one or more proteins (Fig. 24) (100, 101).

Signature sequences also provide evidence that within monoderm prokaryotes, "high-G+C" gram-positive bacteria are the closest relatives of gram-negative bacteria. Phylogenies and signature sequences in a number of proteins provide evidence that all diderm prokaryotes (gram-negative bacteria) are monophyletic and had a common ancestor. The species of the genera *Deinococcus* and *Thermus* are indicated to be interme-

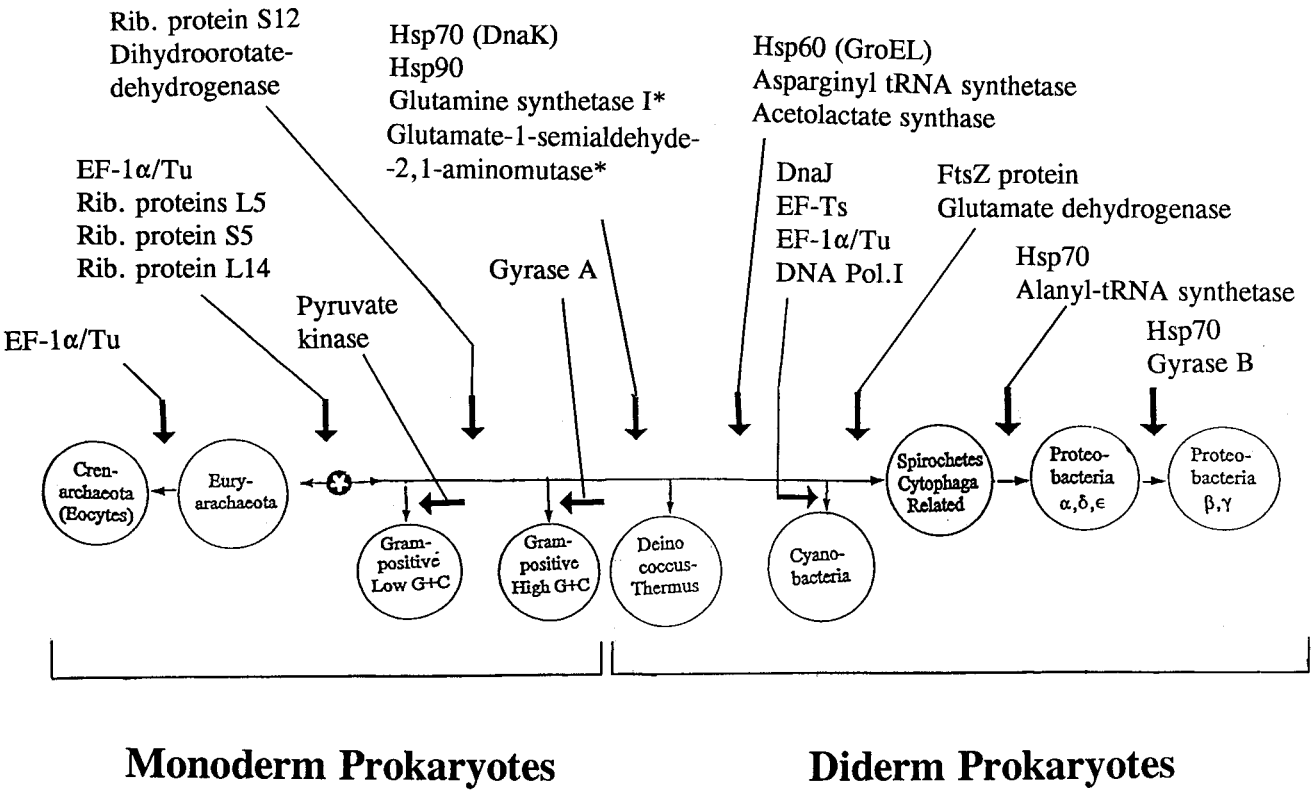


FIG. 24. Evolutionary relationships within prokaryotes as deduced from signature sequences in various proteins. Although, due to ease of presentation, this figure depicts archaebacteria as distinct from other prokaryotes, the alternate view where archaebacteria are derived from gram-positive bacteria (Fig. 23) is favored based on the available evidence. Beginning with the universal ancestor (✪), the order of evolution of different prokaryotic groups as deduced from signature sequences in different proteins is as shown in this diagram. The asterisks on certain proteins indicate that the timing when these signature sequences were introduced may change with sequence information from additional bacterial phyla. The branching order of various eubacterial groups is consistent with the detailed phylogenies based on Hsp70 and GroEL sequences (Fig. 8 and 11).

diate in this transition by both their phenotypic characteristics and molecular sequence data. The molecular phylogenies and phenotypic characteristics again show good agreement in this regard. The sequence data indicate that in the transition from a monoderm prokaryote to a diderm cell organization (i.e., gram-negative bacteria), the outer membrane developed first, followed by changes in the cell wall and other characteristics.

Within diderm prokaryotes, signature sequences and phylogenies based on several genes and proteins provide evidence that cyanobacteria are one of the earliest lineages. This group of bacteria which are capable of carrying out oxygenic photosynthesis, had a profound influence on the environment. Based on the oxygen requirement and oxygen sensitivity of different biochemical reaction, Schopf (208) has concluded that cyanobacteria occupy a middle ground between the anaerobes and the fully aerobic bacteria and eukaryotes, suggesting that this group originated during a time of fluctuating oxygen concentration. The development of oxygenic photosynthesis by cyanobacteria and the consequent release of oxygen into the atmosphere was very likely the key event that changed the environment from anaerobic to aerobic. Based on the geological and mineral evidence of the major episode of sedimentation of dissolved iron from oceans (i.e., banded iron formations), which is believed to have resulted from the release of oxygen by the earliest oxygenic photosynthetic organisms, the time when such organisms first evolved can be estimated to be between 2.0 and 2.5 billion years ago (132, 141, 208).

As mentioned above, although cyanobacteria are physiologically and phylogenetically distinct from the *Deinococcus* and *Thermus* genera, signature sequences in several proteins (DnaJ, EF-Tu, EF-Ts, and DNA polymerase) indicate that these two groups had a common ancestor exclusive of all other prokaryotes. The presence of unique shared sequence signatures in these two groups, which is inconsistent with the morphological features and phylogenetic relationship deduced from other sequences, is very likely a result of lateral gene transfer between the two groups. Similar to the situation encountered in the relationship between archaebacteria and gram-positive bacteria, it is unclear whether the gene transfer took place from cyanobacteria to *Deinococcus* and *Thermus* or vice versa. It would be helpful in this context to know if the transferred genes offered any selective advantages to the recipient organisms. While such information is lacking, one can speculate that the oxygen released by cyanobacteria was highly toxic to the *Deinococcus* and *Thermus* group of species and that transfer of selected genes from cyanobacteria (which have developed a mechanism to protect themselves from oxygen) provided a means to survive in the oxygen-containing atmosphere.

Following the evolution of cyanobacteria, a number of other groups of diderm bacteria, namely, cytophagas, spirochetes, planctomycetes, and green sulfur bacteria, evolved. Because of the paucity of sequence information on these bacterial phyla, no unique signature sequences that can distinguish these groups of bacteria from other diderm prokaryotes have so far been identified. The phylogenetic relationships and the relative branching orders of these phyla in most phylogenies, including

rRNA (184, 250), Hsp70 (Fig. 8) (103, 108) and GroEL (Fig. 11) (96, 246), are not resolved. However, these groups of prokaryotes consistently branch in between cyanobacteria and proteobacteria (Fig. 8 and 11). The placement of these bacterial phyla in a group between cyanobacteria and proteobacteria-1 (the alpha, delta, and epsilon subdivisions) can be confidently made based on the signature sequences in the FtsZ and GDH proteins (Fig. 15) on the one hand and Hsp70 and alanyl-tRNA synthetase on the other (Fig. 19). Although at present I have placed all these bacteria in a single group, it is likely that as further sequence information becomes available, additional signature sequences that make clear distinction between these groups and clarify the evolutionary relationships between these phyla will be identified.

Signature sequences in proteins also define and provide distinction between two different groups corresponding to proteobacteria. One group consists of the alpha, delta, and epsilon subdivisions, whereas the other consists of the beta and gamma subdivisions. The association of *Thermomicrobium* with the first group is surprising, and it remains to be determined whether other members of the green nonsulfur group (*Herpetosiphon* and *Chloroflexus*) also show such a relationship. The group consisting of the beta and gamma proteobacteria, based upon signature sequences in different proteins, appears to have evolved most recently among all the various prokaryotic groups or divisions. In earlier work, specific amino acid substitutions in Hsp60 protein that distinguish the alpha proteobacteria from other subdivisions have also been described (96). It is expected that additional signature sequences providing further distinctions between proteobacterial groups will be uncovered in future studies.

The evolutionary picture reconstructed here based upon signature sequences in different proteins (Fig. 24) is in accordance with the phylogenies derived from a number of highly conserved proteins (Fig. 8 and 11).

EVOLUTIONARY RELATIONSHIP BETWEEN EUKARYOTES AND PROKARYOTES

Based on the purported evolutionary relationships among the prokaryotic organisms just considered, we can now ask how the eukaryotic organisms are related to the prokaryotes. The evolutionary relationship between prokaryotes and eukaryotes has been studied based on a large number of sequences and other characteristics (21, 249a, 263), and I do not plan to provide an exhaustive list of these studies or characteristics. Instead, my objective here is to determine what kind of proposal or model for the origin of eukaryotic cells can best account for most of the available molecular sequence data as well as other relevant information.

Some Critical Assumptions in Studying Prokaryote-Eukaryote Relationships

As discussed above (see "Current evolutionary perspective"), the three-domain proposal postulates that the eukaryotes and archaebacteria had a common ancestor exclusive of any eubacteria (258), in other words, that the nuclear genome of the ancestral eukaryotic cell (exclusive of organellar genes) directly descended from an archaebacterial cell. A specific relationship of the eukaryotes to archaebacteria, as suggested in this proposal, is based on the rooting derived from the duplicated gene sequences for EF-1α/Tu and EF-2/G (and the α and β subunits of F- and V-ATPases) (81, 126). Although the subsequent discovery of V-ATPases in bacteria and an F-ATPase in an archaebacterium have called into question the rooting based on ATPase sequences (69, 118), the rooting based on EF-1α/Tu and EF-2/G proteins is widely accepted and continues to have a major influence on the field (7, 126). A similar rooting of the universal tree between archaebacteria and eubacteria has been derived based on homologous aminoacyl (isoleucyl, valyl-, and leucyl) tRNA synthetase sequences (20).

However, the question of the root of the universal tree is a hotly debated and very contentious issue (49, 53, 69, 129, 205). Depending upon the protein sequences that are considered, different types of relationships among the three primary groups (archaebacteria, eubacteria, and eukaryotes) could be observed. These include (i) all three groups equidistant from each other, (ii) archaebacteria and bacteria closely related to each other compared to the eukaryotic homologs, (iii) archaebacteria as specific relatives of eukaryotes and eubacteria distantly related to them, and (iv) a specific relationship of eukaryotic homologs to eubacteria as compared to the archaebacteria (49, 53, 69, 85, 197a, 205). Since all of the indicated relationships are strongly supported for different proteins, rationally it is difficult to choose among them unless it is postulated that extensive lateral gene transfer occurred between species to support one relationship in preference to the other.

This controversy, in my view, has stemmed in large part from a very basic and profound assumption that the eukaryotic cells have evolved from prokaryotes by normal evolutionary mechanisms (mutations, recombination, etc.). However, as emphasized by many prominent biologists (28, 34, 46, 159, 168, 173, 237), the transition from prokaryotes to eukaryotes represents a major discontinuity in the evolutionary history. If the prokaryote-to-eukaryote transition came about by normal evolutionary mechanisms, then given the enormity of the structural and molecular differences between these two cell types, this transformation must have occurred over a very long period involving numerous intermediate species, each developing limited selective advantages and evolving certain eukaryotic characteristics. However, there is no evidence (living or fossil) for the existence of any such "intermediate" organisms, despite the great diversity of the prokaryotic and eukaryotic organisms that preceded or followed this major change. However, if the transition from prokaryote to eukaryote did not come about by a normal evolutionary mechanism but instead resulted from some unusual event such as fusion and integration of genomes from different prokaryotes, any attempt to root the eukaryotic tree based on any one particular gene or even a set of genes (e.g., the EF-1α/Tu, EF-2/G, and aminoacyl-tRNA synthetases genes) will provide information about the origin of that gene (or sets of genes) and not of the eukaryotic cell. In view of these considerations, the proper approach to understanding the origin of the eukaryotic cell from prokaryotic ancestors, in my view, would be to examine the relationship of different eukaryotic genes to prokaryotic homologs without any prior assumptions and then to suggest a model which is consistent with most of the data. This is the approach I have followed in this review.

Most Genes for the Information Transfer Processes Are Derived from Archaebacteria

For a number of the gene and protein sequences originally studied, namely, EF-1α/Tu, EF-2/G, RNA polymerase II and III subunits and F- and V-type ATPases, the eukaryotic homologs exhibited greater similarity to archaebacteria than to eubacteria (81, 126, 196). A close and specific relationship of archaebacteria to the eukaryotic homologs is also supported by a number of other genes e.g., ribosomal proteins, DNA poly-

merase B, TATA binding proteins, transcription factors IIB and IIIB, TCP-1 chaperone (54, 96, 98, 134, 158, 203, 213), most of which are involved in aspects of transcription and translation. In the past 2 or 3 years, due to the complete sequencing of the genomes of a number of bacterial and archaebacterial species and a eukaryotic species, *Saccharomyces cerevisiae* (15, 26, 45, 66, 72, 73, 80, 119, 128, 138, 147, 215, 242), a much larger database has become available to examine the relationships among species in the three domains. Detailed analyses of the archaebacterium *Methanococcus jannaschii* sequences indicate that 44% of its gene products showed a closer relationship to eubacteria whereas about 13% of the proteins showed a closer relationship to the eukaryotic homologs (144). The rest of the proteins showed approximately the same level of similarity to the eubacterial and eukaryotic homologs. An important understanding that resulted from such analyses is that the vast majority of genes for which the archaebacterial homologs exhibited greater similarity to eukaryotes were related to the information transfer processes such as replication, transcription, and translation (9, 47, 54, 144, 158, 183, 197). In fact, for many genes involved in DNA replication and transcription, no eubacterial homologs have been found (54, 158, 183, 197). Thus, in terms of their informational transfer machineries "archaea look very eukaryotic" (54).

A close and specific relationship between archaebacteria and eukaryotes for the information transfer processes is also readily apparent from signature sequences in many proteins. EF-1α/Tu (Fig. 7a) and ribosomal proteins L5 (Fig. 7b) and S5 (Fig. 7c) provide examples where all archaebacterial and eukaryotic homologs contain prominent shared signature sequences not found in any eubacterial homologs. The protein EF-1α/Tu contains another important signature sequence identified by Rivera and Lake (198). This signature sequence consists of a 7-aa insert that is uniquely present in the eocytes or *Crenarchaeota* group of archaebacteria and all eukaryotic homologs but is not found in the *Euryarchaeota* division of archaebacteria (Fig. 21). This signature provides evidence that within archaebacteria, the eocyte archaebacteria are the closest relatives of eukaryotes (Fig. 25) (198). A specific relationship of the eocyte archaebacteria to eukaryotic homologs is also strongly supported by the detailed phylogenetic studies based on EF-1α/Tu sequences (7, 21, 112). Based on the above observations, it is now indisputable that many of the eukaryotic nuclear genes, particularly those related to the information transfer machinery, are of archaebacterial origin. In relation to the origin of eukaryotic cells, the key question now becomes whether all of the eukaryotic nuclear-cytosolic genome (i.e., exclusive of organelles) are derived from archaebacteria or whether other groups of prokaryotes (i.e., eubacteria) also made significant contributions.

Hsp70 Provides the Clearest Example of the Contribution of Eubacteria to the Nuclear-Cytosolic Genome

The question of establishment of any eubacterial contribution to the eukaryotic nuclear-cytosolic genome is far more difficult than connecting archaebacteria and eukaryotes. (Note that the term "nuclear-cytosolic" as used in this review refers to those genes and proteins which originated with the formation of the ancestral eukaryotic cell.) The main difficulty lies in the fact that in contrast to archaebacteria, which have contributed only to the nuclear genome, two classes of eukaryotic cell organelle genomes, mitochondria and plastids, were derived from eubacteria in later endosymbiotic acquisitions (90, 92, 159). Most organellar genes were later transferred to the nucleus. Thus, eukaryotes often have multiple homologs of proteins with sequence similarity to eubacteria. For most sequences in the databases, information that distinguishes nuclear-cytosolic genes from organellar homologs is lacking. Thus, in many cases it is difficult to know whether a given eukaryotic homolog corresponds to an organellar gene or to a nuclear-cytosolic gene product. The presence of multiple genes inside eukaryotes also raises the possibility that genes for some presumed nuclear-cytosolic proteins are in fact derived from organellar genes by means of horizontal transfer followed by divergence (135, 165). Furthermore, many eukaryotes harbor bacterial endosymbionts, and some of the eubacterial genes could be derived from them by horizontal transfers. Thus, it has proven difficult to establish whether a given eubacterial gene present in eukaryotic cells is of nuclear-cytosolic origin or is derived from other sources. However, in recent years, the enlarged sequence database, in conjunction with extensive characterization of many eukaryotic protein families at the molecular, biochemical, subcellular localization, and phylogenetic levels, has provided specific examples where these problems can be clearly resolved.

The Hsp70 protein, discussed above, provides the best-studied examples of such proteins (17, 99, 102, 108). The Hsp70 homologs have been sequenced and characterized from a broad range of prokaryotic and eukaryotic organisms covering the entire evolutionary spectrum. In eukaryotic cells, specific Hsp70 homologs are found in various cellular compartments, including the cytosol, endoplasmic reticulum (ER), mitochondria, and chloroplasts (17, 41, 102, 108). The homologs present in different compartments are well characterized both biochemically and by cellular localization studies (2, 214, 221). Most importantly, global alignment of Hsp70 sequences from prokaryotic, eukaryotic, and organellar sources shows that the different types of Hsp70 homologs are readily and unambiguously distinguished from each other based on specific signature sequences (99, 102, 105, 108). The eukaryotic nuclear-cytosolic Hsp70s contain a large number of unique amino acid substitutions and sequence signatures not found in any of the prokaryotic or organellar homologs. Figure 26 gives an excerpt from Hsp70 alignment showing some of the important sequence signatures. As seen in this figure, all the eukaryotic cytosolic and ER homologs, including those from the earliest-diverging eukaryotic lineage such as *Giardia*, contain two signature sequences (a 4-aa deletion marked ② and a 1-aa insert marked ③) that are not present in any prokaryotic or organellar Hsp70s. Since these signatures are not present in any mitochondrial, hydrogenosomal (from *Trichomona vaginalis*), or prokaryotic homologs, they could not have been derived from the latter groups by means of horizontal gene transfer. These signatures are thus uniquely eukaryotic, and they were probably introduced into the common ancestor of eukaryotic cells at the time of its origin. The homologs containing these sequence signatures thus are nuclear-cytosolic in origin.

The inference from the above signature sequences that the eukaryotic nuclear-cytosolic homologs are altogether distinct from organellar homologs is strongly reinforced by the phylogenetic analysis based on Hsp70 sequences (Fig. 27). In phylogenetic trees based on Hsp70 sequences, the nuclear-cytosolic homologs consistently form a distinct monophyletic clade (100% of the time by different phylogenetic methods) branching within gram-negative bacteria but showing no relationship to the organellar homologs (i.e., mitochondria, hydrogenosome, or chloroplasts) (56, 102–104, 108). If the mitochondrial and nuclear-cytosolic homologs were paralogous sequences that originated by a gene duplication event and subsequent divergence, one would expect them to form distinct but related

clades. Since this is not observed in any phylogenetic trees based on Hsp70 sequences, such a possibility is considered highly unlikely. In contrast to the cytosolic homologs, the mitochondrial and hydrogenosomal Hsp70s grouped within the same clade, showing an expected close relationship to the alpha proteobacteria (Fig. 27) (25, 78). The homologs from these two organelles also shared a number of unique amino acid substitutions with the alpha proteobacteria (signatures marked E2 in Fig. 28) (25, 78), providing additional evidence of their origin from this group of prokaryotes. The absence of these signatures in the nuclear-cytosolic homologs provides further evidence that they have originated independently of these groups. In phylogenetic trees based on Hsp70, the chloroplast homologs showed the expected strong affinity to cyanobacteria, supporting their origin from this group of prokaryotes (90, 92, 159, 162, 170a, 246a).

Having presented evidence that the cytosolic homologs of Hsp70s are of nuclear-cytosolic origin which originated independently of organellar homologs and that their sequence characteristics and phylogeny cannot be accounted for by lateral gene transfers between species, the question of which group of prokaryotes contributed them now arises. As seen in Fig. 26, all of the nuclear-cytosolic homologs contain the large insert in their N-terminal quadrant (box marked ①), which is a defining characteristic of gram-negative bacteria (i.e., diderm insert). As mentioned above, this insert is not found in any homolog from archaebacteria or gram-positive bacteria. The presence of this shared insert in all nuclear-cytosolic homologs and gram-negative bacteria provides strong evidence that these eukaryotic homologs have originated from a gram-negative bacterium rather than from archaebacteria. This inference is strongly supported by phylogenetic analysis based on Hsp70 sequences (Fig. 27) (85, 102, 108).

The lineage of the gram-negative bacteria that contributed the eukaryotic nuclear-cytosolic homologs has not been identified up to now. However, based upon signature sequences in the Hsp70 family of proteins, it is now possible to infer that the gram-negative bacterium from which these are derived was a member of the proteobacteria-1 (which includes members of the alpha, delta, and epsilon subdivisions as well as *Thermomicrobium*) group (Fig. 28). This inference is based upon the facts that all nuclear-cytosolic homologs of Hsp70 contain a 2-aa insert (signature P1) that is present in different proteobacteria (as well as the green nonsulfur bacterium *T. roseum*) but do not contain the 4-aa insert (signature P2) which is specific for the beta and gamma proteobacteria (i.e., proteobacteria-2). The eukaryotic nuclear-cytosolic homologs also share two additional sequence signatures (specific amino acid substitutions marked E1 in Fig. 28) with the proteobacteria, indicating their origin from this group. However, it should be emphasized that the proteobacteria in general, and the proteobacteria-1 group as defined here in particular, is one of the most diverse and complex assemblages of prokaryotes (177). Although this group includes the alpha proteobacteria, from which mitochondria are derived (57, 90, 162, 261), it also includes numerous other genera of very divergent prokaryotes such as myxobacteria, sulfur- and sulfate-reducing bacteria, helicobacteria, *Campylobacter*, and green nonsulfur bacteria. Thus, a specific relationship of the eukaryotic nuclear-cytosolic homologs to the proteobacteria-1 group should not be construed as evidence that they are derived from the same lineage that gave rise to mitochondrial homologs. As pointed out above, the eukaryotic nuclear-cytosolic homologs do not branch with the mitochondrial homologs in any of the phylogenetic trees and are distinguished from these homologs by numerous sequence features (Fig. 26 and 28) (17, 56, 85, 99, 102–105).

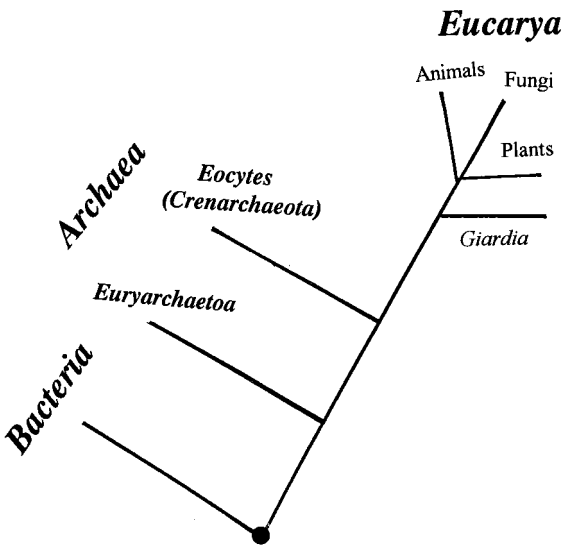


FIG. 25. The eocyte version of the archaebacterial tree based on signature sequence in the EF-1α/Tu protein sequences, as suggested by Rivera and Lake (198). This tree indicates that the ancestral eukaryotic cell has directly descended from within the archaebacterial lineage, with eocyte archaebacteria as its closest relatives.

The Eukaryotic Nuclear Genome Is a Chimera of Genes Derived from Archaebacteria and Gram-Negative Bacteria

In addition to Hsp70, many other examples of proteins where the eukaryotic nuclear-cytosolic homologs are derived from eubacteria and not archaebacteria are observed. (i) In the Hsp90 protein family, which carry out a molecular chaperone function in cells (193), the eukaryotic nuclear-cytosolic homologs (including from *Giardia lamblia* [unpublished results]) and those from gram-negative bacteria contain a 5-aa deletion not present in low-G+C or high-G+C gram-positive bacteria (Fig. 29a). The Hsp90 homologs from eukaryotic cells have been well characterized, and thus far no organellar homolog of Hsp90 has been identified in any species, including the genomic sequence of *Saccharomyces cerevisiae* (80). It is also of interest that thus far no Hsp90 homolog has been found in any archaebacteria, including the three completely sequenced genomes from *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, and *Archaeoglobus fulgidus* (26, 138, 215). Therefore, it is very likely that Hsp90 homologs do not exist in archaebacteria. For the sake of argument, even if such homologs were to be found in other archaebacteria, then based on our understanding of the relationships within prokaryotes (see "Evolutionary relationships among prokaryotes"), such homologs should exhibit closer relationship to the gram-positive bacteria than to the gram-negative bacteria. From the above observations and considerations, it should be clear that similar to Hsp70, the nuclear-cytosolic homologs of Hsp90 are not of archaebacterial origin but instead are derived from gram-negative eubacteria. (ii) In IMP dehydrogenase, all eubacterial and eukaryotic homologs contain a 2-aa conserved indel not present in any archaebacteria (Fig. 29b). Likewise, in adenylosuccinate synthetase, a 2-aa insert is present in various archaebacteria but not in any of the eubacterial or eukaryotic homologs (Fig. 29c). (iii) A number of other proteins where signature sequences uniquely shared between eukaryotic homologs and certain groups of gram-negative bacteria have been observed: glutamate-1-semialdehyde 2,1-aminomutase (Fig. 9b), alanyl-tRNA synthetase (Fig. 19b), and the FGA-

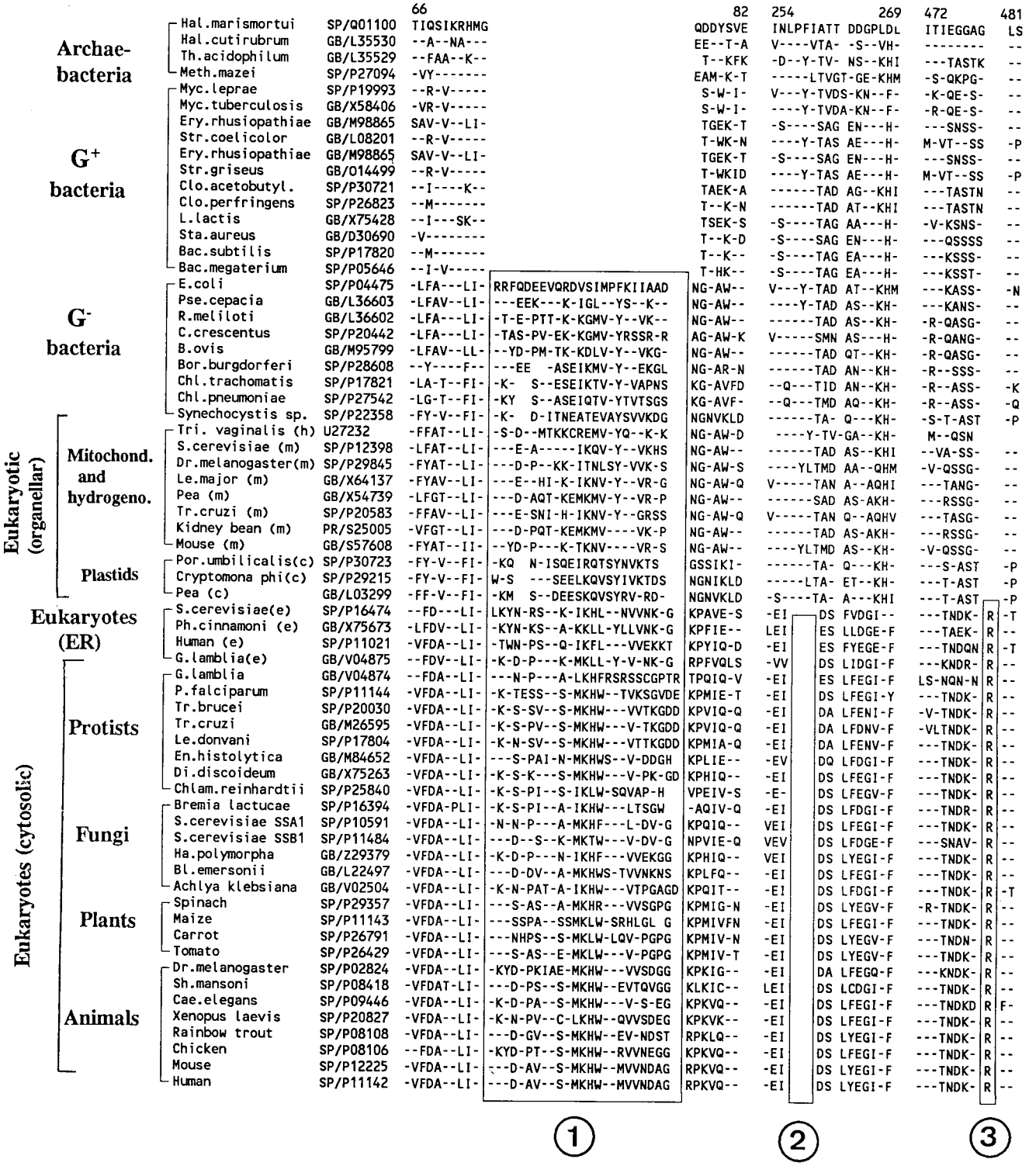


FIG. 26. Excerpt from the Hsp70 sequence alignment showing some of the important sequence signatures (boxed regions) distinguishing eukaryotic nuclear-cytosolic homologs from prokaryotic and organellar homologs. G^- and G^+ refer to gram-negative bacteria and gram-positive bacteria, respectively. The boxed region marked ① shows the diderm insert in the N-terminal quadrant common to all eukaryotic homologs and gram-negative bacteria. The signatures marked ② and ③ identify two indels that distinguish eukaryotic nuclear-cytosolic homologs from all organellar and prokaryotic homologs. Other prokaryotic homologs not included in this alignment (e.g., some shown in Fig. 3) also contained the indicated signature sequences. Not all signature sequences of the above kinds are shown. The notation (e) in parentheses identifies ER Hsp70 homologs. Mitochond. and hydrogeno. refer to mitochondria and hydrogenosome homologs.

RAT protein (105). These signatures provide evidence that the eukaryotic homologs of these proteins are of eubacterial rather than archaebacterial origin.

In a recent study, Feng et al. (63) reported that of the 34 protein families that they examined, for 17 a closer relationship of the eukaryotic homologs to eubacteria was observed. In contrast, only 8 of the 34 proteins indicated a eukaryote-archaebacteria relationship. In addition to these results, our BLAST searches of the proteins encoded by the *Haemophilus influenzae* genome (66) have identified a number of proteins for which the eubacterial and eukaryotic homologs are present but no related protein has thus far been found in archaebacteria (Table 3). Although some of these proteins may correspond to organellar homologs, or for some archaebacteria homologs showing closer affinity may be found in the future, it is likely that for many of these proteins the nuclear-cytosolic homologs are again derived from gram-negative bacteria rather than archaebacteria. Another striking characteristic of eukaryotic cells not explained by an archaebacterial origin is their membrane lipid composition (156). All eukaryotic cell membranes contain ester-linked fatty acid lipids like eubacteria rather than the ether-linked lipids that define archaebacteria (127, 258). Thus, the eukaryotic cell membranes are of eubacterial rather than archaebacterial origin. It should be clear from the above examples that eubacteria have also made significant contribution to the eukaryotic nuclear-cytosolic genes. Hence, the premise that archaebacteria and the ancestor eukaryotic cell had a common ancestor exclusive of eubacteria is doubtful.

Upon examination of different gene and protein phylogenies, where the relationships between the prokaryotic and eukaryotic homologs have been studied, one finds that these generally fall into one of the following three groups: those favoring an archaebacterial-eukaryote clade, those supporting a gram-negative-eukaryote clade, and a third equivocal group where the phylogenies are unable to support or refute any specific relationship between prokaryotes and eukaryotes (49, 53, 69, 85, 126, 197a, 205, 258). The overall inference from these phylogenies that the eukaryotic homologs in different cases show greater similarity to either archaebacteria or gram-negative bacteria is thus consistent with the signature sequences in various proteins described here.

The question now arises of how we can explain the mutually discordant histories of different eukaryotic nuclear genes, where some genes (particularly those related to the information transfer processes) are clearly derived from archaebacteria whereas many others show a close affinity to the gram-negative bacteria (and it is unlikely that they are derived from organellar homologs). These results cannot be explained or accounted for by the three-domain proposal (258) or the eocyte tree (149, 198), which posits that the eukaryotic cell and archaebacteria (or eocytes) had a common ancestor exclusive of any eubacteria. Likewise, other proposals for the origin of eukaryotic cells, including evolution of eukaryotic cells from a transient intermediate between archaebacteria and gram-positive bacteria (30), evolution of eukaryotic cells by engulfment of an archaebacterium by a hypothetical protoeukaryotic lineage that contained RNA-based metabolism (109, 217), origin of eukaryotic cell nucleocytoplasm from an archaebacterium such as *Thermoplams acidophilum* (212) and undulipodia (i.e., motility components such as microtubules) from a spirochete (161, 162), and evolution of eukaryotic cell from a hypothetical prokaryotic lineage that somehow developed phagocytic capacity (46), also cannot account for or explain these results.

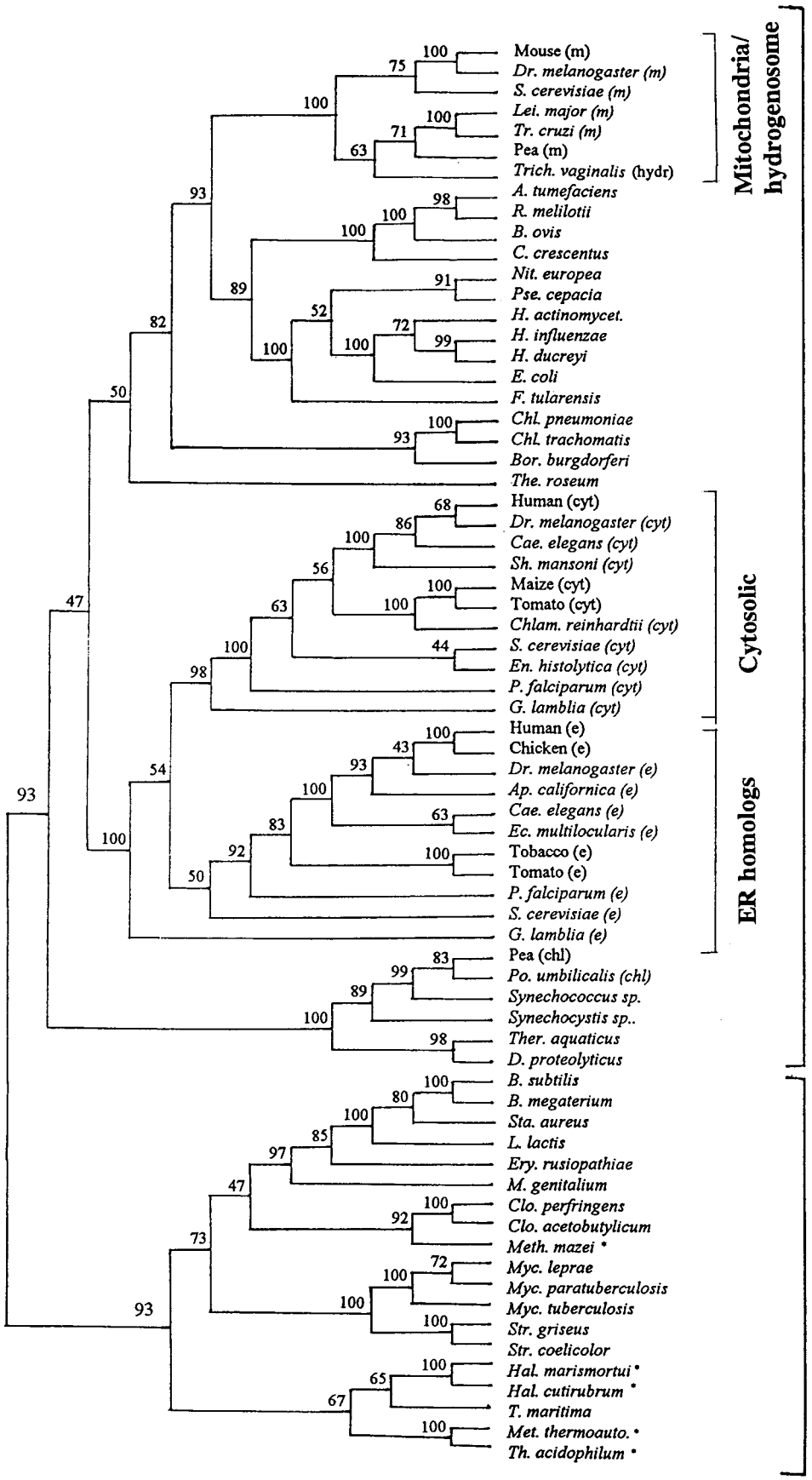
To explain the global phylogenies of eukaryotic nuclear-cytosolic genes and proteins, we have proposed that the ancestral eukaryotic cell, rather than originating from any one particular group of prokaryotes, evolved by means of a unique fusion event between an archaebacterium and a gram-negative bacterium (Fig. 1) (85, 99, 102, 105, 108, 125). The observation of Rivera and Lake (198) on EF-1α sequences suggests that the archaebacterial partner in this fusion was an eocyte or *Crenarchaeota* archaebacterium, and the signature sequences in Hsp70 now provide evidence that the eubacterial partner belonged to the proteobacteria-1 lineage (Fig. 28). At an early stage after the suggested fusion, an assortment or selection of genes from the two fusion partners occurred, during which most of the genes for information transfer such as replication, transcription, and translation were retained from the archaebacterial partner whereas many of the genes for other components and functions such as membrane lipids, Hsp70, Hsp90, adenylosuccinate synthetase, IMP dehydrogenase, FGARAT, and alanyl-tRNA synthetase, were kept from the gram-negative bacterium. The ancestral eukaryotic cell is thus a true chimera that retained and integrated different characteristics from each of the prokaryotic parents (85, 99, 102, 105, 108).

It should be mentioned that the chimeric origin of the ancestral eukaryotic cell by a fusion between an archaebacterium and a eubacterium was first proposed by Zillig (263) to account for the observation that while the large subunits of eukaryotic RNA polymerase II and III exhibited greater similarity to archaebacteria, RNA polymerase I appeared more closely related to eubacteria (196). However, in the later work of Zillig and coworkers, this chimeric model was not favored (139, 158). Lake et al. (154) also mentioned the possibility that the eukaryotic cell nucleus was derived via endosymbiotic capture of an archaebacterium by a eubacteria, but this possibility was not supported in later work (149, 150, 198). However, more recently, based upon the increasing evidence pointing to the contribution of both archaebacteria and gram-negative bacteria to the eukaryotic nuclear genome, many investigators have supported a chimeric origin of the ancestral eukaryotic cell (52, 63, 156, 163, 166, 218).

Origin of the Nucleus and Endoplasmic Reticulum

The key defining characteristics of all eukaryotic cells is the presence of a membrane-bounded nucleus, and hence any insight relating to the origin of the nucleus or the events which accompanied its formation should be of central importance in understanding the origin of the eukaryotic cell. Important insight into this regard is provided by the proteins which are found in the ER. Since the ER is contiguous and forms the nuclear envelope in eukaryotic cells (3, 162), its evolution most probably took place in concert with the nucleus. Thus, the origin of the proteins found in the ER and their evolutionary relationship to other prokaryotic and eukaryotic proteins is critical in understanding the origin of the nucleus. For a number of proteins (Hsp70 and Hsp90) which function as molecular chaperones in the transport of other "passenger proteins" across membranes (40, 190), distinct homologs are found in the ER and cytosolic compartments in all eukaryotic species examined (17, 97, 102, 180).

In our earlier work (102), both ER and cytosolic homologs for Hsp70s were cloned from *Giardia lamblia*, which is one of the earliest-branching eukaryotic lineages (32, 102, 112, 219, 234). The two types of homologs could be readily distinguished based on a number of different sequence features including the N-terminal ER targeting sequence and a C-terminal ER retention signal. The cloning of an ER Hsp70 homolog from *G. lamblia* provided the first strong molecular evidence that the ER originated very early in the eukaryotic cell (102). Direct



evidence for the presence of an ER and of a complex endomembrane system in *G. lamblia* has now been obtained by immunoelectron microscopy with antibodies to the ER Hsp70 (221). Based upon their signature sequences, both the ER and cytosolic Hsp70s from all eukaryotic organisms are shown to be of nuclear-cytosolic origin and are derived from gram-negative bacteria (Fig. 26). Although both cytosolic and ER Hsp70s contain numerous unique shared sequence signatures not found in any prokaryotic or organellar homologs, phylogenetic analyses of these sequences show that they form paralogous gene families (Fig. 27) that evolved by a gene duplication event very early in the evolution of eukaryotic cells (102).

Phylogenetic studies with another molecular chaperone protein, Hsp90, also clearly indicate that the ER and cytosolic forms of this protein form paralogous gene families that resulted from an ancient gene duplication event (Fig. 30) (97). The ER homologs of Hsp90, in addition to their characteristic N-terminal ER-targeting sequence and C-terminal ER retention sequence (97), can be distinguished from the cytosolic homologs by a 2-aa insert present in them (signature ② in Fig. 31). Figure 31 also shows another signature sequence in Hsp90 (marked ①), which distinguishes all eukaryotic homologs from the prokaryotic homologs. Similar to the Hsp70 protein, this signature was again probably introduced into the common ancestor of eukaryotic cells at the time of its origin. In addition to Hsp70 and Hsp90, preliminary evidence also exists that the cytosolic and ER forms of another molecular chaperone protein, DnaJ/Hsp40, also resulted from a gene duplication event at a very early stage in eukaryotic cell history (reference 27 and unpublished results).

The question should be asked why these molecular chaperones are present in the ER and why duplication of their genes accompanied, or was necessary for, the origin and evolution of the ER (nucleus). As mentioned above, one of the main functions of these molecular chaperone proteins is that they facilitate protein transport across intracellular membranes (40, 190, 193). To account for these observations as well as the chimeric nature of eukaryotic nuclear genes, we have suggested that the ancestral eukaryotic cell originated by a unique fusion event between a gram-negative bacterium and an eocyte archaebacterium (99, 102, 105). Although the details of this fusion event are not clear, it is postulated to be distinct from a normal endosymbiotic event (154, 156, 210), where the guest species retains its structural identity at least in a vestigial form. In a simplistic scenario (Fig. 32), a gram-negative eubacterium, probably lacking a cell wall, developed a symbiotic relationship with an archaebacterium. This symbiotic relationship led to the loss of the outer membrane from the gram-negative partner, which no longer needed it to shield itself from antibiotics in the external environment. The loss or extensive divergence of many genes which were no longer essential under these conditions from the two partners also took place under these conditions. The eukaryote-specific signature sequences present in many genes were also probably introduced at this early stage. Over time, the bacterial partner developed numerous membrane infolds that completely surrounded the archaebacterium. The detachment of these membrane infolds from the bacterium eventually led to the creation of the ER, which surrounded the archaebacterium. The membrane of the archaebacterial partner became redundant under these conditions and was eventually lost (Fig. 32). The formation of the nuclear envelope and ER by detachment of membrane infolds would create a new compartment in the cell, which had to communicate (i.e., import and export proteins and other molecules) with the rest of the cell. Therefore, the formation of this compartment was either accompanied or, more likely, preceded by duplication of the genes for the chaperone proteins (Hsp70, Hsp90, DnaJ, etc.), which are essential for this purpose (97, 102). Subsequently, the genome of the eubacterial partner was transferred to the newly formed nucleus, leading to a complete integration of the two parental cells types and the creation of a new cell: the common ancestor of all eukaryotes (97, 99, 102). It should be mentioned that unlike other endosymbiotic events leading to the origins of mitochondria and plastids, which have resulted in the formation of cells with "host plus endosymbiont" phenotypes, the primary fusion event postulated here involved complete integration and loss of identity of the two fusion partners, creating a new cell which was very different from a simple combination of the two fusion partners, i.e., "archaebacterium plus eubacterium."

The phylogenies and signature sequences in different genes and proteins provide strong evidence that the postulated fusion event that gave rise to the ancestral eukaryotic cell was unique and that a successful fusion between prokaryotic parents that gave rise to the eukaryotic cell took place only once in the history of life (99). The evidence for this is derived from signature sequences in a number of proteins, namely, Hsp70 (Fig. 26), Hsp90 (Fig. 31), and glucose-6-phosphate transaminase (Fig. 33), which are unique to all eukaryotic nuclear-cytosolic homologs but are not found in any prokaryotic or organellar homologs. These eukaryotic specific signatures were probably introduced into the common ancestor of eukaryotic cells at the time of its formation and then passed on to all descendants. The presence of these unique signature sequences provides strong evidence that all extant eukaryotic species are monophyletic (99, 102, 105, 250).

The origin of eukaryotic cell by a unique fusion event involving two different groups of prokaryotes, as suggested here, is preferable to the three-domain model, or a number of other proposals for the origin of the eukaryotic cell, for the following reasons. (i) In contrast to the three-domain proposal, which accounts for only some of the gene phylogenies, the chimeric model is the most parsimonious way to explain all of the gene and protein sequence data. (ii) Unlike a number of earlier proposals which postulate the origin of eukaryotic cell from some hypothetical lineages possessing unique characteristics (30, 46, 109, 217), the present model indicates that the ancestral eukaryotic cell was derived from prokaryotic parents related to the extant lineages. (iii) It readily explains why certain characteristics of eukaryotic cells are similar to archaebacteria (e.g., components of transcription and translation machinery) while others are clearly derived from eubacteria (e.g., ester-linked straight-chain membrane lipids, fatty acids, Hsp70, Hsp90, and adenylosuccinate synthetase). (iv) It provides a

FIG. 27. A consensus neighbor-joining tree based on Hsp70 sequences (bootstrapped 100 times) showing the relationship between prokaryotic and various eukaryotic homologs. The tree is based on 531 aligned amino acid positions. The main points to be noted are as follows: mitochondrial and chloroplasts homologs show a specific relationship to the α proteobacteria and cyanobacteria, respectively; the hydrogenosome homolog from *Trichomonas* branches with the mitochondrial clade; the eukaryotic nuclear-cytosolic homologs form a distinct clade within gram-negative bacteria unrelated to the organellar homologs; the ER and cytosolic homologs form paralogous gene families; and archaebacterial homologs (marked with asterisks) show polyphyletic branching within gram-positive bacteria. In the tree shown, only a small number of divergent eukaryotic homologs are included. However, inclusion of additional eukaryotic homologs does not alter the phylogenetic relationship shown here (unpublished results).

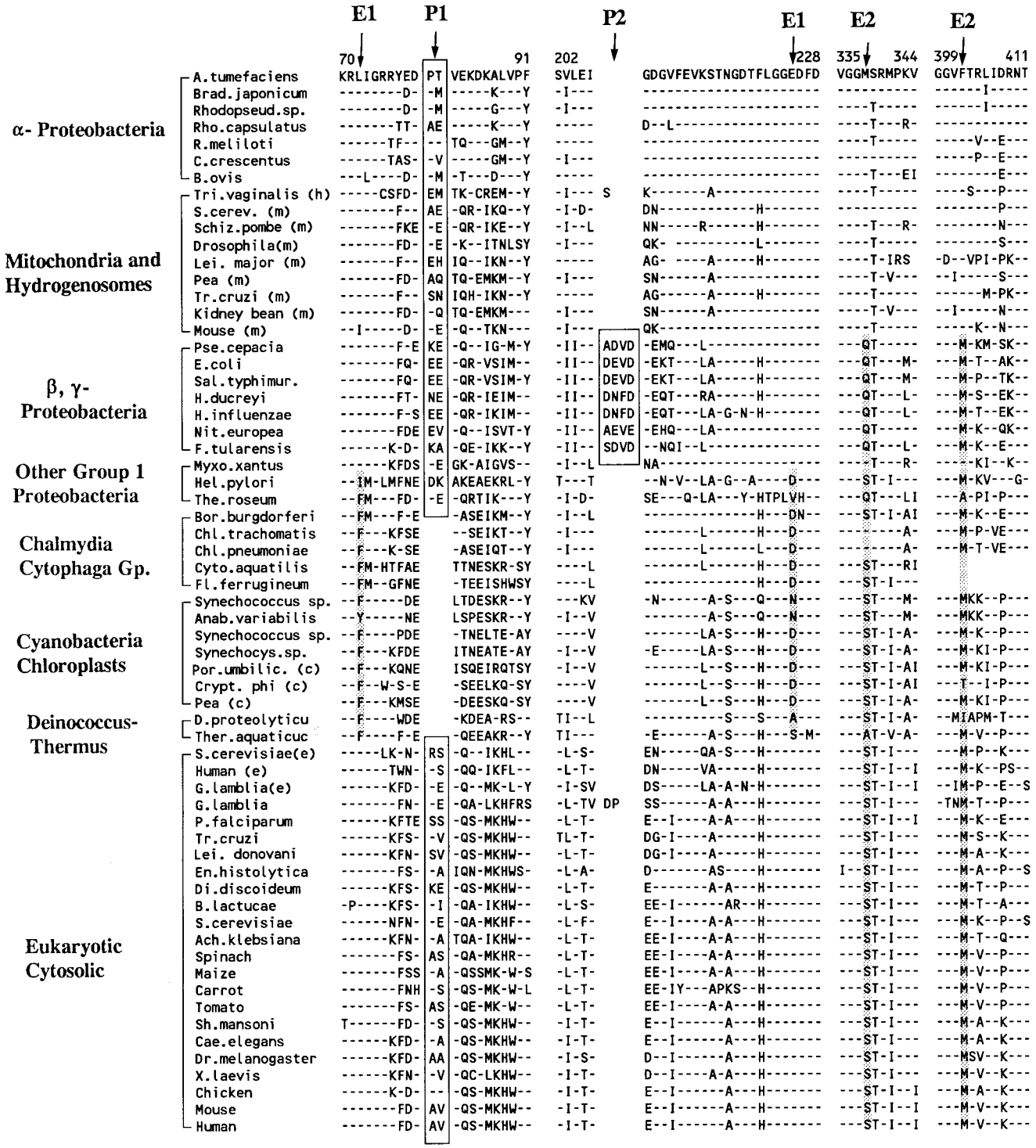


FIG. 28. Signature sequences (boxed and shaded) in the Hsp70 protein showing the relationship of eukaryotic cytosolic homologs to proteobacteria-1 group (alpha, delta, and epsilon subdivisions as well as *Thermomicrobium roseum*). The homologs from various prokaryotic phyla as well as different eukaryotic homologs are identified. The notations (m), (c), (e), and (h) denote mitochondrial, chloroplast, ER, and hydrogenosome homologs, respectively. The signatures P1 and P2 identify sequences that distinguish between proteobacteria-1 and -2. The presence in all nuclear-cytosolic homologs of the 2-aa proteobacteria-1 signature but not the 4-aa proteobacteria-2 signature provides evidence that these homologs are derived from a member of the proteobacteria-1 group. The signatures marked E1 are also common to proteobacteria-1 and proteobacteria-2 as well as eukaryotic cytosolic Hsp70s, supporting the above inference. The signatures E2 identify two substitutions that are present in all members of the alpha proteobacteria as well as mitochondrial and hydrogenosome homologs but absent in other groups of proteobacteria and eukaryotic cytosolic homologs. These signature suggest that the eukaryotic nuclear-cytosolic homologs have originated independently of mitochondria and hydrogenosomes.

plausible explanation for the origin of the eukaryotic cell nucleus and endomembrane systems. (v) Although the enormous structural differences between the eukaryotic and prokaryotic cell types (169) and the absence of any intermediates in this transition, cannot be readily explained by normal evolutionary mechanisms, this major evolutionary discontinuity can be explained by an origin of the eukaryotic cell by fusion of two different groups of prokaryotes. (vi) Doolittle and coworkers (51, 63) have inferred that the eukaryotic species diverged from either archaebacteria or eubacteria about 2 Ga ago based on genetic distances between protein sequences. Although these estimates involve many assumptions (84) (see "Molecular phylogenies: assumptions, limitations, and pitfalls"), the inferences derived are consistent with the present model.

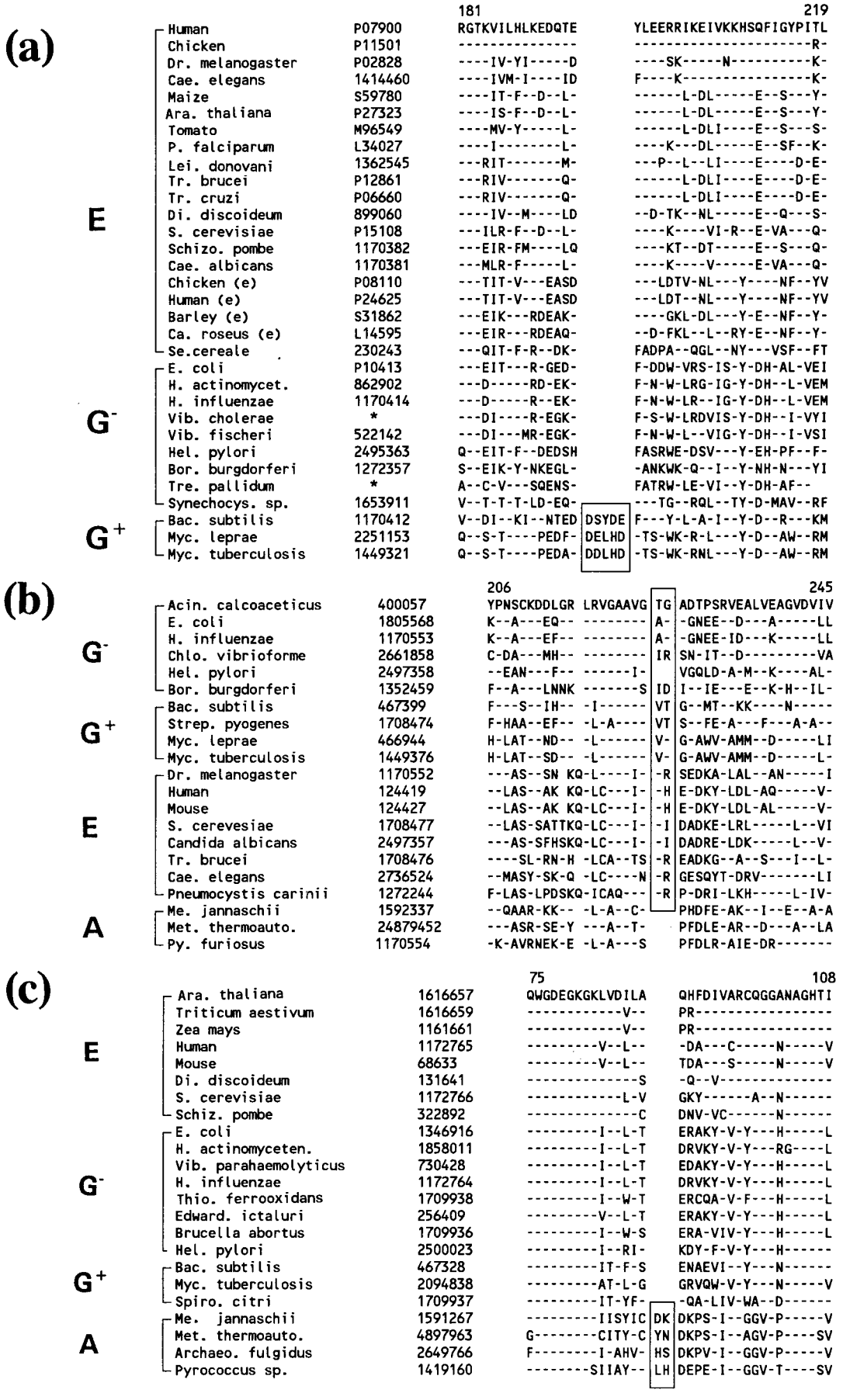
Did Mitochondria and the First Eukaryotic Cell Originate from the Same Fusion Event?

In the past, it has been generally accepted that the ancestral eukaryotic cell lacked mitochondria, which were acquired in a later endosymbiotic event (90, 92, 159, 162, 258). However, the recent finding of certain glycolytic and fermentation enzymes, i.e., glyceraldehyde-3-phosphate dehydrogenase, triosephosphate isomerase, pyruvate:ferredoxin oxidoreductase, ferredoxin, and alcohol dehydrogenase E, and other mitochondrion-specific proteins (e.g., Hsp60 and mitHsp70) in a number of protist phyla, namely, Parabasala (e.g., *Trichomonas vaginalis*), Archamoebae (e.g., *Entamoeba*, *Pelomyxa*), Microsporidia (*Vairimorpha nectarix*), and Diplomonands (*Giardia*), which were previously thought to lack mitochondria, has suggested that the mitochondrial endosymbiosis occurred much earlier than was previously suspected (25, 35, 52, 78, 113a, 120, 123, 135, 165, 199–201, 218). Based on these studies, while the exact time when mitochondria originated will no doubt be moved considerably earlier, the interpretation of these results concerning the origin of eukaryotic cells requires caution. The phylogenies based on glycolytic and fermentation enzymes are ambiguous. For glyceraldehyde-3-phosphate dehydrogenase, multiple homologs are present in both prokaryotic and eukaryotic species and their relationships to each other is not clear (116, 165, 201). The phylogenies of pyruvate:ferredoxin oxidoreductase, ferredoxin, and alcohol dehydrogenase E have led Rosenthal et al. (201) to conclude that the eukaryotic genes for these proteins in early-branching protists were derived from bacteria by means of horizontal gene transfers.

More credible evidence for the presence of mitochondrial genes in early-branching protists comes from the study of the heat shock molecular chaperone proteins Hsp60 (or Cpn60) and Hsp70 (25, 35, 78, 120, 123, 199, 200, 218), where the phylogenies and the relationship between different homologs are well understood (17, 96, 98, 105, 108, 246). In eukaryotes, Hsp60 genes are derived primarily from organellar genomes (i.e., mitochondria and chloroplasts) and no close homologs of nuclear-cytosolic origin are known (96, 98). Thus, the presence of any Hsp60 gene in a species is generally taken as evidence that it once contained mitochondria. However, it is important to point out that mitochondrial Hsp60 possesses no unique sequence characteristic, except for the presence of a N-terminal mitochondrial targeting presequence (MTP), by which it could be distinguished from bacterial homologs. In contrast to Hsp60, all eukaryotes contain a number of distinct Hsp70 homologs. The mitochondrial Hsp70 homologs, in species which contain mitochondria, are clearly distinguished from nuclear-cytosolic homologs by signature sequences and phylogenetic branching patterns (Fig. 26 and 27), but they are indistinguishable from the alpha proteobacterial homologs, except for the presence of a N-terminal MTP (214). Thus, the main basis for concluding that a given Hsp60 or Hsp70 homolog, from a species lacking mitochondria, is of mitochondrial origin is based on three lines of evidence: (i) localization of the homolog to a subcellular compartment that may be related to mitochondria (i.e., hydrogenosomes), (ii) the presence of characteristic MTP sequences found in mitochondrial homologs, and (iii) branching of the homologs with the mitochondrial clade in phylogenetic trees. Of these three lines of evidences, in my view the first two are more reliable indicators of mitochondrial origin. Since many protist species, including *Giardia lamblia*, harbor intracellular bacterial symbionts and/or surface-attached bacteria (1, 160), some of which could be derived from the same group of prokaryotes as mitochondria, based on the branching pattern of the homolog with the mitochondrial clade alone, the possibility that the observed gene is either a bacterial contaminant or derived from bacteria via horizontal gene transfer cannot be excluded.

Examined in this light, there is good evidence that Hsp60 or Hsp70 genes identified in *Trichomonas vaginalis* are of mitochondrial origin. These genes are localized in hydrogenosomes which, based on biochemical criteria, are related to mitochondria (172); some of these genes contain targeting sequence similar to those found in mitochondria; and they branch consistently with mitochondria in independent studies (25, 78, 199). The homologs for these proteins in *Entamoeba histolytica* and *Vairimorpha nectarix* are also probably of mitochondrial origin, since in addition to their branching with the mitochondrial clade, they contain MTP-like sequences (35, 120). However, evidence for the ancestral presence of mitochondria in diplomonads from studies on *G. lamblia* (200, 218), which constitutes one of the earliest-branching lineages in many gene phylogenies (32, 102, 113, 219, 234), must be viewed with caution. Evidence for the presence of mitochondria in this protist is based mainly on the branching of Hsp60 homolog with the mitochondrial clade (200). The cloned gene contains no N-terminal MTP sequence characteristic of mitochondria or other related organelles. In this context, it should be pointed out that the presence of an Hsp60-related protein in *Giardia* was first suggested in our work based on cross-reactivity of Hsp60 antibodies to a giardial protein (222). However, our cloning studies with *Giardia* (unpublished results) resulted in the isolation of a novel Hsp60 gene that has all the characteristics of a bacterial rather than a mitochondrial gene: (i) it lacks any upstream targeting sequence characteristic of organellar genes, and (ii) it hybridized to *Giardia* cultures grown under standard conditions but showed no hybridization when the cells were grown in the presence of antibiotics such as streptomycin. It should be emphasized that in these studies no bacterial contaminant could be detected by light or electron microscopic investigation or by staining with Hoechst 33258, indicating the cryptic nature of these bacteria (reference 222 and unpublished results). These observations emphasize the need for caution in interpreting the results of finding mitochondrion-like genes in the earliest-branching eukaryotic lineages for the origin of eukaryotic cells.

The question may now be considered whether the primary fusion event that led to the origin of the eukaryotic cell was identical to or distinct from the one that gave rise to mitochondria (52, 129, 166). In view of the recent findings, it is clear that the endosymbiotic event leading to the acquisition of mitochondria took place much earlier than was previously believed (29, 32, 218). However, the available data in my view still strongly indicate that mitochondrial endosymbiosis was distinct from the primary fusion event that gave rise to the ancestral eukaryotic cell. Some key observations which support this con-



tention are as follows. (i) The mitochondria, like later endosymbionts (plastids), have retained most of the structural and functional characteristics of the prokaryotic parent from which they evolved, including their distinct information transfer machinery. There appears to be no direct contribution from archaebacteria to the mitochondrial function. In contrast to the distinctly eubacterial nature of mitochondria and the genes encoding various mitochondrial proteins, the eukaryotic cell and nuclear genome are totally distinct from mitochondria and represent a true integration of different characteristics from both archaebacterial and eubacterial partners, which lost their identity in the process. (ii) For the Hsp70 protein, which represents the best-studied eukaryotic protein family, the mitochondrial and nuclear cytosolic homologs are quite different and show no affinity for each other (102, 103, 108). All of the nuclear-cytosolic homologs of Hsp70 contain a large number of sequence characteristics that are not present in any mitochondrial homologs or alpha proteobacteria. (iii) While the cytosolic and ER-specific Hsp70 have been identified in all eukaryotes, no gene for the mitochondrial Hsp70 has thus far been detected in the earliest-branching eukaryotic lineages such as *Giardia*. (iv) Even if such a gene is identified in *Giardia* in future studies, then to account for the very different sequence characteristics of the mitochondrial and nuclear-cytosolic homologs, one would have to postulate that the endosymbiotic event leading to the formation of the eukaryotic cell was immediately followed by a duplication of genes for the Hsp70 protein and then by extensive divergence of one gene copy corresponding to one of the nuclear-cytosolic homologs. This gene duplication event then needs to be immediately followed by another gene duplication in the earliest eukaryotic ancestor to account for the paralogous families of ER and cytosolic homologs, which are found in all eukaryotic organisms. It would also require that the Hsp70 gene from the archaebacterial host be lost in the earliest eukaryotic ancestor, since no archaebacterium-like Hsp70 is present in any eukaryote. (v) The formation of eukaryotic cell by endosymbiotic capture of a gram-negative bacterium by an archaebacterium does not explain how the eukaryotic cell nucleus and ER were formed and how the membrane of the archaebacterial host was replaced by those of the endosymbiont. The application of Ockham's razor "Non sunt entia multiplicanda practor necessitatum" ("unnecessary assumptions should be avoided in formulating hypotheses") to this problem indicates that it is highly unlikely that the endosymbiotic event which gave rise to mitochondria also resulted in the origin of the ancestral eukaryotic cell.

Lastly, the nature of the selective forces that led to the origin of the eukaryotic cell should be considered. Martin and Muller (166) have recently proposed the hydrogen hypothesis for the formation of the eukaryotic cell, which posits that the eukaryotic cell resulted from a symbiotic association between a hydrogen-dependent archaebacterium (such as a methanogen) and an alpha proteobacterium, which under anaerobic conditions produced molecular hydrogen as a waste product. The driving (or selective) force in this symbiotic association was the dependence of the archaebacterium on the molecular hydrogen produced by the symbiont. Martin and Muller (166), by making different assumptions, have suggested how this single symbiotic event could lead to the origin of the eukaryotic cell, mitochondria, and hydrogenosomes. While the model proposed by Martin and Muller satisfactorily accounts for the origin of mitochondria and hydrogenosomes from the same endosymbiotic event, it does not explain or even consider the phylogeny or sequence characteristics of some of the best-studied eukaryotic protein families which provide the main evidence about the earliest events in the origin of the eukaryotic cell, i.e., formation of the ER and nucleus (see the previous paragraph). In addition to the problems outlined in the previous paragraph, this model for the origin of the eukaryotic cell is inconsistent with the following facts. (i) The endosymbiotic capture of an anaerobic hydrogen-producing bacterium by a strictly anaerobic archaebacterium should produce a cell

TABLE 3. Proteins in the *H. influenzae* genome that are found in both eubacteria and eukaryotes but for which no archaebacterial homologs have been found

Name[a]	Gene identification no.[b]
1-Acyl-*sn*glycerol-3-phosphate acetyltransferase	HI0734
ATP-dependent protease (sms)	HI1597
Acetoacetate CoA-transferase (α subunit)	HI0774
Catalase	HI0928
Dehydroquinase	HI0970
Deoxyribose-phosphate aldolase	HI0047
7,8-dihydro-6-hydroxymethylpterin phosphokinase	HI0064
Dihydrolipoamide acetyltransferase	HI1232
Dihydrolipoamide succinyltransferase	HI1661
Dihydropterate synthase	HI1336
DNA binding protein HU	HI0430
DNA mismatch repair protein (MutL)	HI0667
DNA polymerase I	HI0856
Formyltetrahydrofolate hydrolase	HI1588
Fructose-1,6-bisphosphatase	HI1645
Galactokinase	HI0819
Glucose6-phosphate-1-dehydrogenase	HI0558
GTP cyclohydrolase I	HI1447
Ketoacyl reductase	HI0155
Leukotoxin secretion ATP binding protein	HI1051
5,10-Methylenetetrahydrofolate dyhydrogenase	HI1444
NAD(P) transhydrogenase (α subunit)	HI1362
Peptide chain release factor 1	HI1561
Peptidyl-tRNA hydrolase homolog	HI0394
Phosphatidylserine synthase	HI0425
Phosphatidylserine decarboxylase proenzyme	HI0160
Protein translocase subunit (SecA)	HI0909
Replicative DNA helicase (DnaB)	HI1574
Ribonucleoside-diphosphate reductase	HI1660
Ribosomal protein L9	HI0544
Ribosomal protein L16	HI0784
Ribosomal protein L19	HI0201
Ribosomal protein L20	HI1320
Ribosomal protein S6	HI0547
S-Adenosylmethionine synthetase	HI1172
Single-stranded DNA binding protein	HI0250
Thioredoxin	HI1115
Translation initiation factor IF3	HI1318
Uracil DNA glycosylase	HI0018
Uridine kinase	HI0132

[a] CoA, coenzyme A.
[b] Refers to the gene identification number in the *H. influenzae* genome (66).

FIG. 29. Signature sequences (boxed) in the Hsp90 (a), IMP dehydrogenase (b), adenylosuccinate synthetase (c) proteins showing the relatedness of the eukaryotic cytosolic homologs (E) to eubacteria (G^+ and G^-) rather than archaebacteria (A). For Hsp90, no archaebacterial homolog has been identified in the three genomes that have been completely sequenced (26, 138, 215).

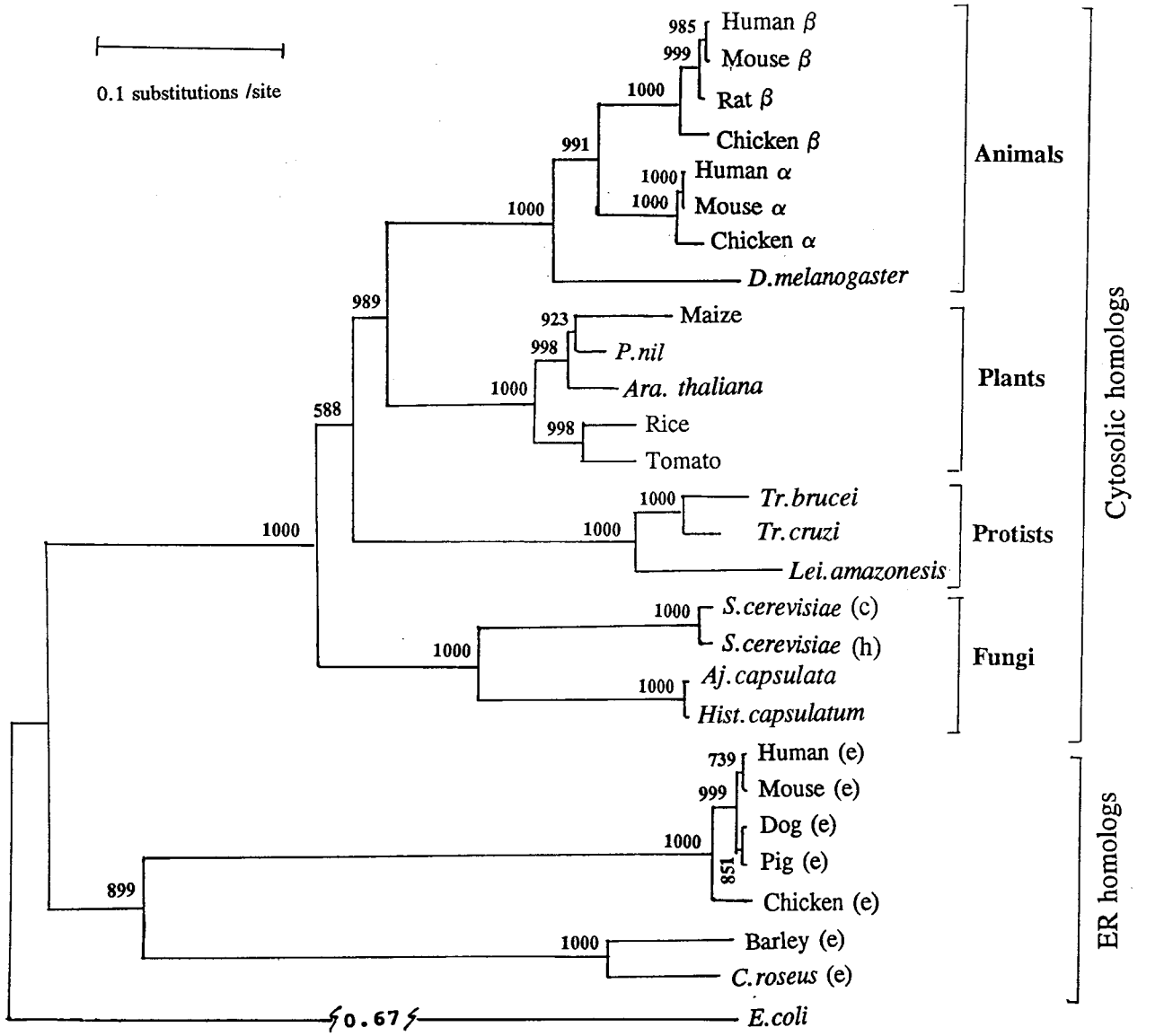


FIG. 30. Neighbor-joining distance tree based on Hsp90 sequences indicating that the cytosolic and ER resident forms of these protein form paralogous gene families, which resulted from a gene duplication event very early in the history of eukaryotic cells. The bootstrap scores out of 1,000 replicates are shown. Reproduced from reference 97 with permission of the publisher.

which should also be a strict anaerobe; however, to my knowledge, no free-living eukaryotic organism is strictly anaerobic. (ii) Since this fusion took place in an oxygenic atmosphere (based upon the evolutionary position of proteobacteria in the prokaryotic lineage [see "Evolutionary relationships among eukaryotes"]), an anaerobic cell will be at a great selective disadvantage under such conditions. (iii) Since endosymbiotic association between hydrogen-producing bacteria and hydrogen-dependent archaebacteria is indicated to be very common, it does not explain the uniqueness of the fusion event. (iv) Molecular sequence data indicate that among archaebacteria, the eocyte group of archaebacteria (i.e., thermoacidophilic) and not methanogens are the closest relatives of eukaryotes (198).

In contrast to the hydrogen hypothesis, which posits hydrogen dependence as the major selective force, I propose that the two major selective forces that had a profound influence in shaping the evolutionary history of life were (i) the antibiotic selection pressure, which probably led to the evolution of both archaebacteria and the diderm prokaryotes (see "Possible selective forces leading to horizontal gene transfers") and (ii) oxygen sensitivity of the organisms when the atmosphere changed from anaerobic to aerobic (208). In my view, a combination of these two selective forces led to the association and ultimate fusion of an antibiotic-resistant archaebacterium with an oxygen-tolerant eubacterium to produce a novel eukaryotic cell which was antibiotic resistant and oxygen tolerant. This scenario explains why, during the gene assortment process in the ancestral eukaryotic cell, most of the genes for the information transfer processes (which provide the main targets for different antibiotics) were retained from the archaebacterial partner whereas a large number of genes for the metabolic processes were acquired from the eubacterial parent. To account for the uniqueness of the fusion event, it is likely that the

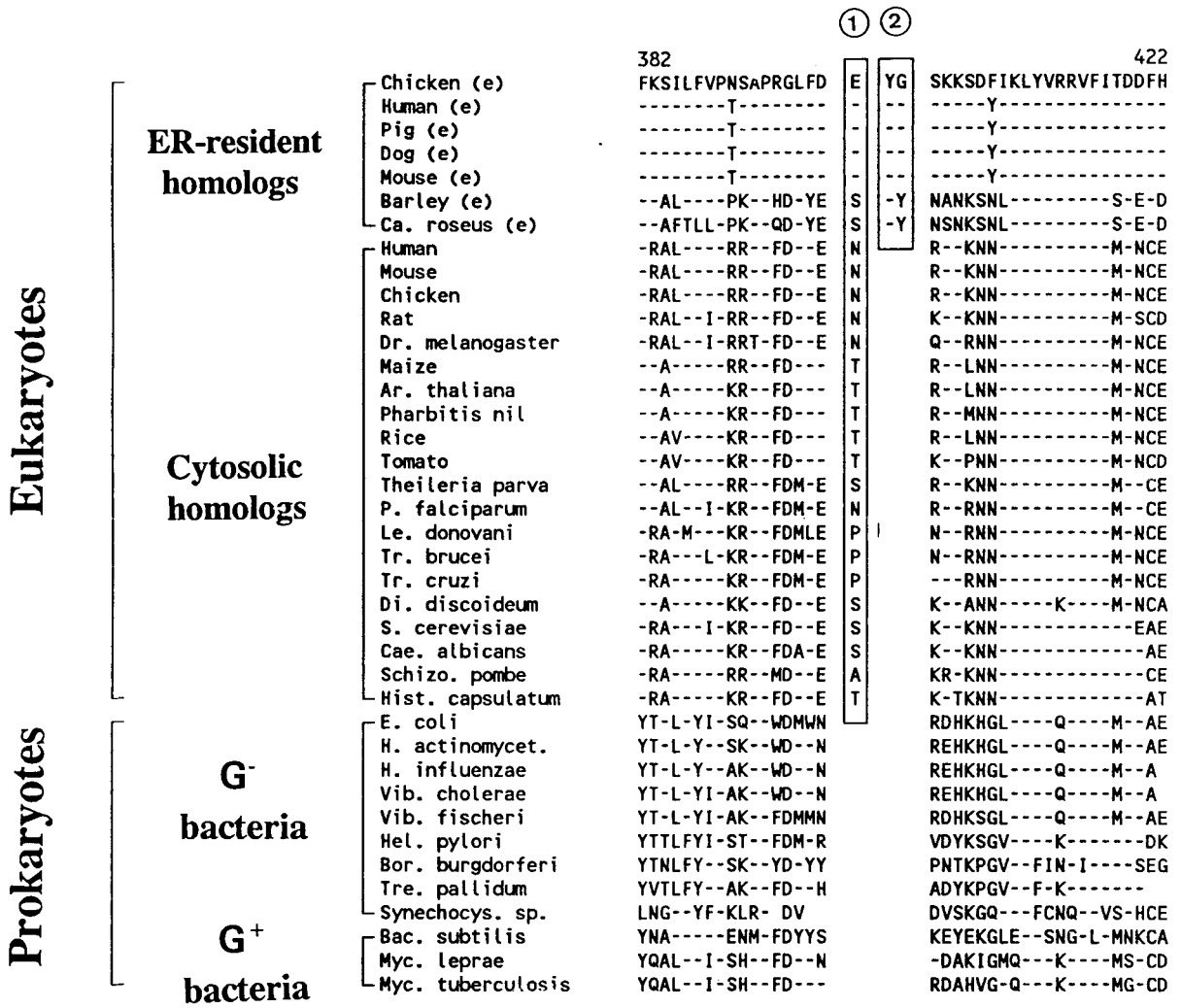


FIG. 31. Signature sequences (boxed) in Hsp90 proteins showing the distinctness of eukaryotic homologs from prokaryotic homologs ① and the distinction between ER homologs and the cytosolic homologs ②.

two groups of prokaryotes came together under a unique set of atmospheric and environmental conditions, which led to an association and selection of the new cell type.

CONCLUDING REMARKS

Signature sequences and phylogenies based on different proteins permit a reconstruction of the basic evolutionary history of prokaryotes involving minimal assumptions. These studies reveal that the evolutionary relationship within the prokaryotic species is a continuum from the earliest-diverging prokaryotes (low-G+C gram-positive bacteria and euryarchaeota archaebacteria) to the most recent groups (beta and gamma proteobacteria), which can be accounted for by normal evolutionary mechanisms. The sequence data on a number of different proteins suggest that the archaebacteria are polyphyletic and are close relatives of gram-positive bacteria. The genes which support a monophyly of archaebacteria are generally those which are targets for the antibiotics produced by gram-positive bacteria. Thus, antibiotic-induced selection pressure may have played an important role in the evolution of archaebacteria, diderm prokaryotes, and eukaryotes. A previously unrecognized and important distinction within prokaryotes, forming the primary taxonomic division within them, which is supported by both molecular sequence data and morphological features, is of the monoderm prokaryotes (*Monodermata*, i.e., those bounded by a single cell membrane) and the diderm prokaryotes (*Didermata*, i.e., those bounded by inner and outer cell membranes defining a periplasmic compartment). In that sense, both archaebacteria and gram-positive bacteria are monoderm prokaryotes, and the distinction between archaebacteria and eubacteria is misplaced. Based on molecular sequences, it is possible to infer that the monoderm prokaryotes are ancestral and the diderm prokaryotes have been derived from them. The signature sequences in different proteins support the division of *Archaebacteria* into two distinct groups (*Euryarchaeota* and *Crenarchaeota*) and of gram-positive bacteria into at least two groups, corresponding to the low-G+C and high-G+C species, of which the high-G+C group is specifically related to the diderm prokaryotes. The *Deinococcus-Thermus* group of species appears to be intermediate in the transition between monoderm (i.e., gram-positive bacteria) and diderm (i.e., gram-negative bacteria) prokaryotes. Within gram-negative bacteria, evolution seems to have proceeded by splitting off new groups in the following order: *Deinococcus*

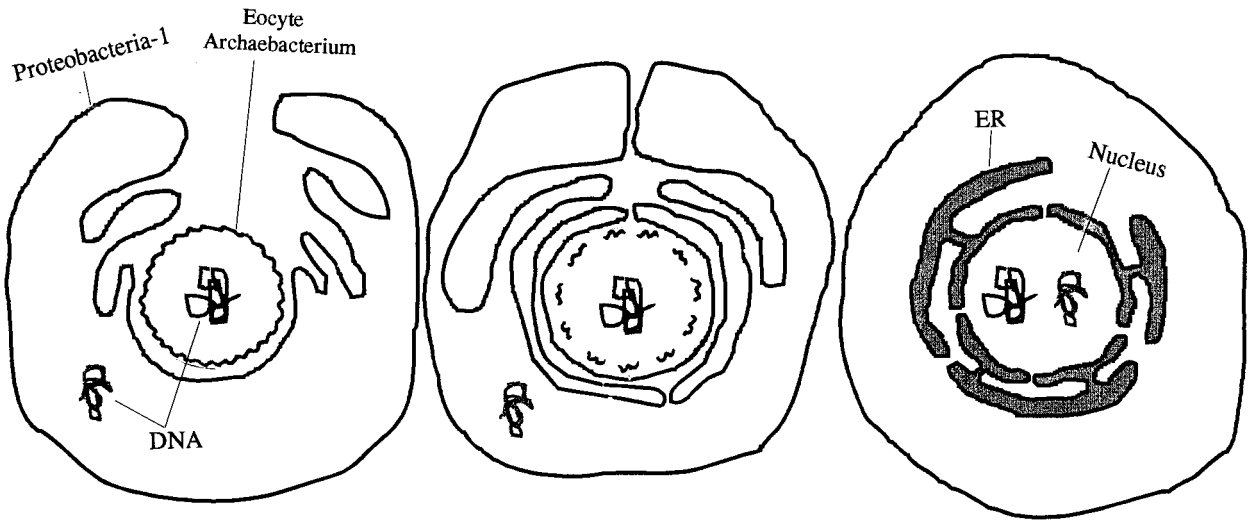


FIG. 32. Origin of the eukaryotic cell nucleus and endomembrane system as per the chimeric model. The key event in the origin of the eukaryotic cell is postulated to be a symbiotic association between a gram-negative eubacterium (from the proteobacteria-1 group) and likely an "eocyte" archaebacterium. This association led to the loss of the outer membrane from the gram-negative bacterium (not shown). As the membrane of the gram-negative bacterium surrounded the eocyte species, the membrane of the latter species, containing ether-linked lipids (wavy line), became redundant and was lost. Eventual separation of the membrane infolds led to the formation of the nuclear envelope and ER. The formation of these new compartments was preceded or accompanied by duplication of the genes for the chaperone proteins (Hsp70, Hsp90, DnaJ, etc.), which are necessary for protein transport and communication within the compartments. The transfer of the genome from the gram-negative eubacterium to the newly formed nucleus and an assortment and integration of genes from the two partners led to the formation of the ancestral eukaryotic cell. Modified and reproduced from reference 105 with permission of the publisher.

and *Thermus* → cyanobacteria → chlamydia, spirochetes and relatives → proteobacteria-1 (includes green nonsulfur bacteria and alpha, delta, and epsilon proteobacteria) → proteobacteria-2 (includes beta and gamma proteobacteria).

The evolutionary history deduced here based on signature sequences in some of the most highly conserved protein sequences in the biota is in contrast to the rather confusing picture that seems to be emerging from other analyses of the completed bacterial genomes (21, 50, 68, 130, 143, 144, 182, 191, 255). However, as has been pointed out (50, 143, 144, 182), of the large number of sequences in individual genomes, many are unique to particular organisms or are found in only closely related species and thus are of limited use for evolutionary studies. Many others show limited sequence conservation, again limiting their usefulness for resolving distant evolutionary relationships. Hence, the number of gene sequences that show a high degree of conservation and can provide reliable evolutionary relationships (unaffected by horizontal gene transfers, etc.) that correlate with the structural and physiological attributes of organisms may turn out to be relatively small. However, the relationships based on these should be consistent with and should help explain other information.

The phylogenies and signature sequences based on a range of proteins also provide evidence that all eukaryotic cells, in-

			1		32
G⁻	E. coli	790167	MCGIVGAIA		QR DVAEILLEGLRRLEYRGYDS
	H. influenzae	169920	-------V-		-- ------IN--H---------
	Thio. ferrooxidans	293522	------GVS		KT -LVPMI----Q---------
	R. legumminosarum	28480	------IVG		HK P-S-R-I-A-G---------
	R. meliloti	128481	------IVG		HQ P-S-R-V-A-EP--------
	Hel. pylori	2314711	------Y-G		DSEKKS V-----KE--------
	Sy. Sp. PCC6803	651800	------Y-G		TQ TAVN--I---E---------
	Ther. aquaticus	184044	------YVG		F- NATDV--D------------
G⁺	Bac. subtilis	161919	------Y-G		-L -AK----K--EK--------
	Myc. tuberculosis	2388656	------YVG		R- PAYVVVMDA---M-------
	Myc. leprae	729589	---L--YVG		-- PACGVVMDA---M-------
A	Met. thermoauto.	726643	-----AC-L		KDGSA-PV---CV----------
	Me. jannaschii	592069	----I-Y-G		ND KAPK---N------------
E	Cae. elegans	149474	----FAYLN	FLAPK	K-SEIVD--VQ--Q-M-------
	Human	544382	----FAYLN	YHVPR	T-REIL-T-IK--Q---------
	Mouse	346130	----FAYLN	YHVPR	T-REIL-T-IK--Q---------
	Candida albicans	707898	----F-YVN	FLVDK	S-GEIIDN-I---Q---------
	S. cerecvisiae	462173	----F-YCN	YLVER	S-GEIIDT-VD--Q---------
	Schiz. pombe	169892	----F-Y-N	YLVER	D-GYILKT-VK--K---------

FIG. 33. Signature sequence in glucose-fructose-6-phosphate transaminase, showing the presence of a unique signature (boxed) in eukaryotic homologs. The eukaryotic homologs for Hsp70 (Fig. 26) and Hsp90 (Fig. 31) also contain several unique sequence signatures not found in any prokaryotic homologs. These signature provides evidence that all of the eukaryotes are derived from a single ancestor and that the postulated fusion event was unique.

cluding amitochondriate and aplastidic cells, received major gene contributions to the nuclear genome from both an archaebacterium (very probably of the eocyte group) and a gram-negative bacterium (related to proteobacteria-1). From these data, it is proposed that in contrast to the basic premise of the three-domain proposal, the ancestral eukaryotic cell never directly descended from archaebacteria but instead was a chimera formed by fusion and integration of the genomes of an archaebacterium and a gram-negative bacterium. The available data indicate that the primary fusion event that gave rise to the ancestral eukaryotic cell was unique and that it was very probably distinct from (and preceded) the one that gave rise to mitochondria and hydrogenosomes. These results provide evidence for an alternative view of the evolutionary relationships among the extant organisms that differs from the three-domain proposal.

ACKNOWLEDGMENTS

I thank Vanessa Johari, Charu Chandrashekhar, and Thuyanh Le for database searches on different proteins. Thanks are also due to R. G. E. Murray, K. B. Freeman, B. J. Soltys, and two anonymous reviewers for their critical reading and many helpful comments on the manuscript. Several helpful suggestions received from Lynn Margulis and R. G. E. Murray concerning taxonomic terms and conventions are also gratefully acknowledged. I am also indebted to B. Singh for his involvement in cloning and sequencing of different genes, which formed the foundation of our work in this area.

The work from my laboratory was supported by a research grant from the Medical Research Council of Canada.

REFERENCES

1. **Adam, R. D.** 1991. The biology of *Giardia* spp. Microbiol. Rev. **55:**706–732.
2. **Ahmad, S., R. Ahuja, T. J. Venner, and R. S. Gupta.** 1990. Identification of a protein altered in mutants resistant to microtubule inhibitors as a member of the major heat shock protein (hsp70) family. Mol. Cell. Biol. **10:**5160–5165.
3. **Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson.** 1994. Molecular biology of the cell. Garland Publishing, Inc., New York, N.Y.
4. **Allsopp, A.** 1969. Phylogenetic relationships of the procaryota and the origin of the eucaryotic cell. New Phytol. **68:**591–612.
5. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
6. **Atlas, R. M.** 1988. Microbiology: fundamentals and applications. Macmillan Publishing Co., New York, N.Y.
7. **Baldauf, S. L., J. D. Palmer, and W. F. Doolittle.** 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA **93:**7749–7754.
8. **Balows, A., H. G. Trüper, M. Dworkin, W. Harder, and K. H. Schleifer.** 1992. The prokaryotes. Springer-Verlag, New York, N.Y.
9. **Belfort, M., and A. Weiner.** 1997. Another bridge between kingdoms: tRNA splicing in archaea and eukaryotes. Cell **89:**1003–1006.
10. **Benachenhou-Lahfa, N., P. Forterre, and B. Labedan.** 1993. Evolution of glutamate dehydrogenase genes: evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. J. Mol. Evol. **36:**335–346.
11. **Beveridge, T. J.** 1990. Mechanism of gram variability in select bacteria. J. Bacteriol. **172:**1609–1620.
12. **Beveridge, T. J., and J. Davies.** 1983. Cellular response of *Bacillus subtilis* and *Escherichia coli* to the Gram stain. J. Bacteriol. **156:**846–858.
13. **Beveridge, T. J., and S. Schultze-Lam.** 1996. The response of selected members of the archaea to the Gram stain. Microbiology **142:**2887–2895.
14. **Black, J. G.** 1993. Microbiology: principles and applications. Prentice-Hall, Inc., Englewood Cliffs, N.J.
15. **Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453–1462.
16. **Bocchetta, M., E. Ceccarelli, R. Creti, A. M. Sanangelantoni, O. Tiboni, and P. Cammarano.** 1995. Arrangement and nucleotide sequence of the gene (fus) encoding elongation factor G (EF-G) from the hyperthermophilic bacterium Aquifex pyrophilus: phylogenetic depth of hyperthermophilic bacteria inferred from analysis of the EF-G/fus sequences. J. Mol. Evol. **41:**803–812.
17. **Boorstein, W. R., T. Ziegelhoffer, and E. A. Craig.** 1994. Molecular evolution of the HSP70 multigene family. J. Mol. Evol. **38:**1–17.
18. **Bork, P., C. Sander, and A. Valencia.** 1992. An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. Proc. Natl. Acad. Sci. USA **89:**7290–7294.

18a. **Brennan, P. J., and H. Nikaido.** 1995. The envelope of mycobacteria. Annu. Rev. Biochem. **64:**29–63.

19. **Brooks, B. W., R. G. E. Murray, J. L. Johnson, E. Stackebrandt, C. R. Woese, and G. E. Fox.** 1980. Red-pigmented micrococci: a basis for taxonomy. Int. J. Syst. Bacteriol. **30:**627–646.
20. **Brown, J. R., and W. F. Doolittle.** 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. USA **92:**2441–2445.
21. **Brown, J. R., and W. F. Doolittle.** 1997. *Archaea* and the prokaryote-to-eukaryote transition. Microbiol. Rev. **61:**456–502.
22. **Brown, J. R., Y. Masuchi, F. T. Robb, and W. F. Doolittle.** 1994. Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J. Mol. Evol. **38:**566–576.
23. **Buchanan, R. E.** 1925. General systematic bacteriology. The Williams & Wilkins Co., Baltimore, Md.
24. **Buchanan, R. E., and N. E. Gibbons.** 1974. Bergey's manual of deteminative bacteriology, The Williams & Wilkins Co., Baltimore, Md.
25. **Bui, E. T., P. J. Bradley, and P. J. Johnson.** 1996. A common evolutionary origin for mitochondria and hydrogenosomes. Proc. Natl. Acad. Sci. USA **93:**9651–9656.
26. **Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, and J. C. Venter.** 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273:**1058–1073.
27. **Bustard, K., and R. S. Gupta.** 1997. The sequences of heat shock protein 40 (DnaJ) homologs provide evidence for a close evolutionary relationship between the *Deinococcus-Thermus* group and cyanobacteria. J. Mol. Evol. **45:**193–205.
28. **Cavalier-Smith, T.** 1986. The kingdoms of organisms. Nature **324:**416–417.
29. **Cavalier-Smith, T.** 1987. Eukaryotes with no mitochondria. Nature **326:**332–333.
30. **Cavalier-Smith, T.** 1987. The origin of eukaryotic and archaebacterial cells. Ann. N. Y. Acad. Sci. **503:**17–54.

30a. **Cavalier-Smith, T.** 1991. The evolution of cells, p. 271–304. *In* S. Osawa and T. Honjo (ed.), Evolution of life. Springer-Verlag, Tokyo, Japan.

31. **Cavalier-Smith, T.** 1992. Origins of secondary metabolism. Ciba Found. Symp. **171:**64–80.
32. **Cavalier-Smith, T., and E. E. Chao.** 1996. Molecular phylogeny of the free-living archezoan Trepomonas agilis and the nature of the first eukaryote. J. Mol. Evol. **43:**551–562.
33. **Cedergren, R., M. W. Gray, Y. Abel, and D. Sankoff.** 1988. The evolutionary relationships among known life forms. J. Mol. Evol. **28:**98–112.
34. **Chatton, E.** 1937. Titres et travaux scientifiques (1906–1937) de Edouard Chatton. E. Sottano, Sete, France.
35. **Clark, C. G., and A. J. Roger.** 1995. Direct evidence for secondary loss of mitochondria in Entamoeba histolytica. Proc. Natl. Acad. Sci. USA **92:**6518–6521.
36. **Cohan, F. M.** 1994. Genetic exchange and evolutionary divergence in prokaryotes. Science **264:**382–388.
37. **Cohan, F. M.** 1996. The role of genetic exchange in bacterial evolution. ASM News **62:**631–636.
38. **Colewell, R. R.** 1970. Polyphasic taxonomy of bacteria, p. 421–436. *In* H. Iizuka and T. Hasegawa (ed.), Culture collections of microorganisms. University of Tokyo Press, Tokyo, Japan.
39. **Counsell, T., and R. G. E. Murray.** 1986. Polar lipid profiles of the genus *Deinococcus*. Int. J. Syst. Bacteriol. **36:**202–206.
40. **Craig, E. A.** 1993. Chaperones: helpers along the pathways to protein folding. Science **260:**1902–1903.
41. **Craig, E. A., B. D. Gambill, and R. J. Nelson.** 1993. Heat shock proteins: molecular chaperones of protein biogenesis. Microbiol. Rev. **57:**402–414.
42. **Creti, R., E. Ceccarelli, M. Bocchetta, A. M. Sanangelantoni, O. Tiboni, P. Palm, and P. Cammarano.** 1994. Evolution of translational elongation factor (EF) sequences: reliability of global phylogenies inferred from EF-1 alpha(Tu) and EF-2(G) proteins. Proc. Natl. Acad. Sci. USA **91:**3255–3259.
43. **Darwin, C.** 1859. The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London, United Kingdom.
44. **Davies, J.** 1994. Inactivation of antibiotics and the dissemination of resistance genes. Science **264:**375–382.
45. **Deckert, G., P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox, D. E. Graham, R. Overbeek, M. A. Snead, M. Keller, M. Aujay, R. Huber, R. A. Feldman, J. M. Short, G. J. Olsen, and R. V. Swanson.** 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature **392:**353–358.
46. **de Duve, C.** 1996. The birth of complex cells. Sci. Am. **274:**50–57.

47. **Dennis, P. P.** 1997. Ancient ciphers: translation in Archaea. Cell **89:**1007–1010.
48. **De Rijk, P., Y. Van de Peer, I. Van den Broeck, and R. De Wachter.** 1995. Evolution according to large ribosomal subunit RNA. J. Mol. Evol. **41:**366–375.
49. **Doolittle, R. F.** 1995. Of archae and eo: what's in a name? Proc. Natl. Acad. Sci. USA **92:**2421–2423.
50. **Doolittle, R. F.** 1998. Microbial genomes opened up. Nature **392:**339–342.
51. **Doolittle, R. F., D. F. Feng, S. Tsang, G. Cho, and E. Little.** 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. Science **271:**470–477.
52. **Doolittle, W. F.** 1998. A paradigm gels shifty. Nature **392:**15–16.
53. **Doolittle, W. F., and J. R. Brown.** 1994. Tempo, mode, the progenote, and the universal root. Proc. Natl. Acad. Sci. USA **91:**6721–6728.
54. **Edgell, D. R., and W. F. Doolittle.** 1997. Archaea and the origin(s) of DNA replication proteins. Cell **89:**995–998.
55. **Eisen, J. A.** 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. J. Mol. Evol. **41:**1105–1123.
56. **Falah, M., and R. S. Gupta.** 1994. Cloning of the *hsp70* (*dnaK*) genes from *Rhizobium meliloti* and *Pseudomonas cepacia*: phylogenetic analyses of mitochondrial origin based on a highly conserved protein sequence. J. Bacteriol. **176:**7748–7753.
57. **Falah, M., and R. S. Gupta.** 1997. Phylogenetic analysis of mycoplasmas based on Hsp70 sequences: cloning of the *dnaK* (*hsp70*) gene region of *Mycoplasma capricolum*. Int. J. Syst. Bacteriol. **47:**38–45.
58. **Felsenstein, J.** 1982. Numerical methods for inferring evolutionary trees. Q. Rev. Biol. **57:**379–404.
59. **Felsenstein, J.** 1985. Confidence limits in phylogenies: an approach using the bootstrap. Evolution **39:**783–791.
60. **Felsenstein, J.** 1988. Phylogenies from molecular sequences: inference and reliability. Ann. Rev. Genet. **22:**521–565.
61. **Felsenstein, J.** 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. **266:**418–427.
62. **Felsenstein, J.** 1997. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. **27:**401–410.
63. **Feng, D. F., G. Cho, and R. F. Doolittle.** 1997. Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA **94:**13028–13033.
64. **Fitch, W. M.** 1997. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool. **20:**406–416.
65. **Fitch, W. M., and E. Margoliash.** 1967. Construction of phylogentic trees: a method based on mutational distances as estimated from cytochrome c sequences is of general applicability. Science **155:**279–284.

65a. **Flaherty, K. M., D. B. McKay, W. Kabsch, and K. C. Holmes.** 1991. Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70-kDa heat shock cognate protein. Proc. Natl. Acad. Sci. USA **88:**5041–5045.

66. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick.** 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269:**496–512.
67. **Forterre, P.** 1996. A hot topic: the origin of hyperthermophiles. Cell **85:**789–792.
68. **Forterre, P.** 1997. Protein versus rRNA: problems in rooting the universal tree of life. ASM News **63:**89–95.
69. **Forterre, P.** 1997. Archaea: what can we learn from their sequences. Curr. Opin. Genet. Dev. **7:**764–770.
70. **Forterre, P., N. Benachenhou-Lahfa, F. Confalonieri, M. Duguet, C. Elie, and B. Labedan.** 1992. The nature of the last universal ancestor and the root of the tree of life, still open questions. Biosystems **28:**15–32.
71. **Fox, G. E., E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen, and C. R. Woese.** 1980. The phylogeny of prokaryotes. Science **209:**457–463.
72. **Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, M. Gwinn, B. Dougherty, J. F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, A. R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. Van Vugt, N. Palmer, M. D. Adams, and J. Gocayne.** 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature **390:**580–586.
73. **Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, and J. M. Kelley.** 1995. The minimal gene complement of *Mycoplasma genitalium*. Science **270:**397–403.
74. **Frye, N.** 1990. Words with power, p. 129. Harcourt Brace Jovanovich, New York, N.Y.
75. **Fuerst, J. A.** 1995. The Planctomycetes: emerging models for microbial ecology, evolution and cell biology. Microbiology **141:**1493–1506.
76. **Galley, K. A., B. Singh, and R. S. Gupta.** 1992. Cloning of HSP70 (dnaK) gene from Clostridium perfringens using a general polymerase chain reaction based approach. Biochim. Biophys. Acta **1130:**203–208.
77. **Galtier, N., and M. Gouy.** 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proc. Natl. Acad. Sci. USA **92:**11317–11321.
78. **Germot, A., H. Philippe, and H. Le Guyader.** 1996. Presence of a mitochondrial-type 70-kDa heat shock protein in Trichomonas vaginalis suggests a very early mitochondrial endosymbiosis in eukaryotes. Proc. Natl. Acad. Sci. USA **93:**14614–14617.
79. **Glasby, J. S.** 1979. Encyclopedia of antibiotics. John Wiley & Sons, Inc., New York, N.Y.
80. **Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver.** 1996. Life with 6000 genes. Science **274:**546, 563–567.
81. **Gogarten, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, and T. Oshima.** 1989. Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA **86:**6661–6665.
82. **Gogarten, J. P., T. Starke, H. Kibak, J. Fishman, and L. Taiz.** 1992. Evolution and isoforms of V-ATPase subunits. J. Exp. Biol. **172:**137–147.
83. **Gogarten-Boekels, M., E. Hilario, and J. P. Gogarten.** 1995. The effects of heavy meteorite bombardment on the early evolution—the emergence of the three domains of life. Origins Life Evol. Biosphere **25:**251–264.
84. **Golding, B.** 1996. Evolution: when was life's first branch point? Curr. Biol. **6:**679–682.
85. **Golding, G. B., and R. S. Gupta.** 1995. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. Evol. **12:**1–6.
86. **Gouy, M., and W. H. Li.** 1989. Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. Nature **339:**145–147.
87. **Gouy, M., and W. H. Li.** 1989. Molecular phylogeny of the kingdoms Animalia, Plantae, and Fungi. Mol. Biol. Evol. **6:**109–122.
88. **Gram, C.** 1884. Ueber die isolierte farbung der Schizomyceten in Schnitt und Trockenpraparaten. Fortschr. Med. **2:**185–189.
89. **Gray, M. W.** 1988. Organelle origins and ribosomal RNA. Biochem. Cell Biol. **66:**325–348.
90. **Gray, M. W.** 1992. The endosymbiont hypothesis revisited. Int. Rev. Cytol. **141:**233–357.
91. **Gray, M. W.** 1996. The third form of life. Nature **383:**299–300.
92. **Gray, M. W., and W. F. Doolittle.** 1982. Has the endosymbiont hypothesis been proven? Microbiol. Rev. **46:**1–42.
93. **Grogan, D. W.** 1996. Exchange of genetic information at extremely high temperatures in the archaeon *Sulfolobus acidocaldarius*. J. Bacteriol. **178:**3207–3211.
94. **Gruber, T. M., and D. A. Bryant.** 1997. Molecular systematic studies of eubacteria, using σ^{70}-type sigma factors of group 1 and group 2. J. Bacteriol. **179:**1734–1747.
95. **Gupta, R. S.** 1990. Sequence and structural homology between a mouse T-complex protein TCP-1 and the 'chaperonin' family of bacterial (GroEL, 60-65 kDa heat shock antigen) and eukaryotic proteins. Biochem. Int. **20:**833–841.
96. **Gupta, R. S.** 1995. Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukaryotic cells. Mol. Microbiol. **15:**1–11.
97. **Gupta, R. S.** 1995. Phylogenetic analysis of the 90 kD heat shock family of protein sequences and an examination of the relationship among animals, plants, and fungi species. Mol. Biol. Evol. **12:**1063–1073.
98. **Gupta, R. S.** 1996. Evolutionary relationships of chaperonins, p. 27–64. *In* R. J. Ellis (ed.), The chaperonins. Academic Press, Inc., New York, N.Y.
99. **Gupta, R. S.** 1997. Protein phylogenies and signature sequences: evolutionary relationships within prokaryotes and between prokaryotes and eukaryotes. Antonie Leeuwenhoek **72:**49–61.
100. **Gupta, R. S.** 1998. Life's third domain (*Archaea*): an established fact or an endangered paradigm? A new proposal for classification of organisms based on protein sequences and cell structure. Theor. Popul. Biol. **54:**91–104.
101. **Gupta, R. S.** 1998. What are archaebacteria: life's third domain or monoderm prokaryotes related to Gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. Mol. Microbiol. **29:**695–708.
102. **Gupta, R. S., K. Aitken, M. Falah, and B. Singh.** 1994. Cloning of *Giardia lamblia* heat shock protein HSP70 homologs: implications regarding origin of eukaryotic cells and of endoplasmic reticulum. Proc. Natl. Acad. Sci. USA **91:**2895–2899.
103. **Gupta, R. S., K. Bustard, M. Falah, and D. Singh.** 1997. Sequencing of heat shock protein 70 (DnaK) homologs from *Deinococcus proteolyticus* and *Thermomicrobium roseum* and their integration in a protein-based phylogeny of prokaryotes. J. Bacteriol. **179:**345–357.
104. **Gupta, R. S., and G. B. Golding.** 1993. Evolution of HSP70 gene and its implications regarding relationships between archaebacteria, eubacteria, and eukaryotes. J. Mol. Evol. **37:**573–582.
105. **Gupta, R. S., and G. B. Golding.** 1996. The origin of the eukaryotic cell. Trends Biochem. Sci. **21:**166–171.
106. **Gupta, R. S., and V. Johari.** 1998. Signature sequences in diverse proteins

provide evidence of a close evolutionary relationship between the *Deinococcus-Thermus* group and cyanobacteria. J. Mol. Evol. **46:**716–720.
107. **Gupta, R. S., and B. Singh.** 1992. Cloning of the *HSP70* gene from *Halobacterium marismortui*: relatedness of archaebacterial *HSP70* to its eubacterial homologs and a model for the evolution of the *HSP70* gene. J. Bacteriol. **174:**4594–4605.
108. **Gupta, R. S., and B. Singh.** 1994. Phylogenetic analysis of 70 kD heat shock protein sequences suggests a chimeric origin for the eukaryotic cell nucleus. Curr. Biol. **4:**1104–1114.
109. **Hartman, H.** 1984. The origin of the eukaryotic cell. Speculations Sci. Technol. **7:**77–81.
110. **Hasegawa, M., and M. Fujiwara.** 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mol. Phylogenet. Evol. **2:**1–5.
111. **Hasegawa, M., and T. Hashimoto.** 1993. Ribosomal RNA trees misleading? Nature **361:**23.
112. **Hashimoto, T., and M. Hasegawa.** 1996. Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1alpha/Tu and 2/G. Adv. Biophys. **32:**73–120.
113. **Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto, and M. Hasegawa.** 1994. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, Giardia lamblia. Mol. Biol. Evol. **11:**65–71.
113a. **Hashimoto, T., L. B. Sanchez, T. Shirakura, M. Muller, and M. Hasegawa.** 1998. Secondary loss of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. Proc. Natl. Acad. Sci. USA **95:**6860–6865.
114. **Hensel, R., P. Zwicki, S. Fabry, J. Lang, and P. Palm.** 1997. Sequence comparison of glyceraldehyde-3-phosphate dehydrogenases from the three urkingdoms: evolutionary implications. Can. J. Microbiol. **35:**81–85.
115. **Hensell, R., W. Demharter, O. Kandler, R. M. Kroppenstedt, and E. Stacke brandt.** 1986. Chemotaxonomic and molecular-genetic studies of the genus *Thermus*: evidence for a phylogenetic relationship of *Thermus aquaticus* and *Thermus ruber* to the genus *Deinococcus*. Int. J. Syst. Bacteriol. **36:**444–453.
116. **Henze, K., A. Badr, M. Wettern, R. Cerff, and W. Martin.** 1995. A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. Proc. Natl. Acad. Sci. USA **92:**9122–9126.
117. **Higgins, D. G., and P. M. Sharp.** 1988. CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. Gene **73:**237–244.
118. **Hilario, E., and J. P. Gogarten.** 1993. Horizontal transfer of ATPase genes—the tree of life becomes a net of life. Biosystems **31:**111–119.
119. **Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann.** 1996. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res. **24:**4420–4449.
120. **Hirt, R. P., B. Healy, C. R. Vossbrinck, E. U. Canning, and T. M. Embley.** 1997. A mitochondrial *Hsp70* orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. Curr. Biol. **7:**995–998.
121. **Holt, J. G., N. R. Krieg, P. H. A. Sneath, J. T. Staley, and S. T. Williams.** 1994. Bergey's manual of determinative bacteriology, 9th ed. The Williams & Wilkins Co., Baltimore, Md.
122. **Hori, H., and S. Osawa.** 1987. Origin and evolution of Organisms as deduced from 5S ribosomal RNA sequences. Mol. Biol. Evol. **4:**445–472.
123. **Horner, D. S., R. P. Hirt, S. Kilvington, D. Lloyd, and T. M. Embley.** 1996. Molecular data suggest an early acquisition of the mitochondrion endosymbiont. Proc. R. Soc. London Ser. B **263:**1053–1059.
124. **Inouye, M.** 1979. What is the outer membrane? p. 1–12. *In* M. Inouye (ed.), Bacterial outer membranes: biogenesis and functions. John Wiley & Sons, Inc., New York, N.Y.
125. **Irwin, D. M.** 1994. Molecular evolution. Who are the parents of eukaryotes? Curr. Biol. **4:**1115–1117.
126. **Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata.** 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86:**9355–9359.
127. **Kandler, O., and H. Konig.** 1993. Cell envelopes of archaea: structure and chemistry, p. 223–259. *In* M. Kates, D. J. Kushner, and A. T. Matheson (ed.), The biochemistry of Archaea (Archaebacteria). Elsevier Science Publishers B.V. New York, N.Y.
128. **Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, and S. Tabata.** 1996. Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. **3:**109–136.
129. **Karlin, S., J. Mrazek, and A. M. Campbell.** 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. **179:**3899–3913.
130. **Karlin, S., and J. Mrázek.** 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. USA **94:**10227–10232.
131. **Karlin, S., G. M. Weinstock, and V. Brendel.** 1995. Bacterial classifications derived from RecA protein sequence comparisons. J. Bacteriol. **177:**6881–6893.
132. **Kasting, J. F.** 1993. Earth's early atmosphere. Science **259:**920–926.
133. **Kates, M.** 1992. Archaebacterial lipids: structure, biosynthesis and function. Biochem. Soc. Symp. **58:**51–72.
134. **Keeling, P. J., and W. F. Doolittle.** 1995. Archaea: narrowing the gap between prokaryotes and eukaryotes. Proc. Natl. Acad. Sci. USA **92:**5761–5764.
135. **Keeling, P. J., and W. F. Doolittle.** 1997. Evidence that eukaryotic triosephosphate isomerase is of alpha-proteobacterial origin. Proc. Natl. Acad. Sci. USA **94:**1270–1275.
136. **Kimura, M.** 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.
137. **Kishino, H., and M. Hasegawa.** 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. **29:**170–179.
138. **Klenk, H. P., R. A. Clayton, J. F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, D. L. Richardson, A. R. Kerlavage, D. E. Graham, N. C. Kyrpides, R. D. Fleischmann, J. Quackenbush, N. H. Lee, G. G. Sutton, S. Gill, E. F. Kirkness, B. A. Dougherty, K. McKenney, M. D. Adams, and B. Loftus.** 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature **390:**364–370.
139. **Klenk, H. P., and W. Zillig.** 1994. DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. J. Mol. Evol. **38:**420–432.
140. **Kluyver, A. J., and C. B. van Niel.** 1936. Prospects for a natural system of classification of bacteria. Zentbl. Bakteriol. Parasitenkd. Infektionskr. Hyg. Abt. II **94:**369–403.
141. **Knoll, A. H.** 1992. The early evolution of eukaryotes: a geological perspective. Science **256:**622–627.
142. **Kondratieva, E. N., N. Pfennig, and H. G. Trüper.** 1992. The Phototrophic Prokaryotes, p. 312–330. *In* A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K. H. Schleifer (ed.), The prokaryotes, 2nd ed. Springer-Verlag, New York, N.Y.
143. **Koonin, E. V., and M. Y. Galperin.** 1997. Prokaryotic genomes: the emerging paradigm of genome-based microbiology. Curr. Opin. Genet. Dev. **7:**757–763.
144. **Koonin, E. V., A. R. Mushegian, M. Y. Galperin, and D. R. Walker.** 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol. Microbiol. **25:**619–637.
145. **Koonin, E. V., A. R. Mushegian, and K. E. Rudd.** 1996. Sequencing and analysis of bacterial genomes. Curr. Biol. **6:**404–416.
146. **Kristensen, T., R. Lopez, and H. Prydz.** 1992. An estimate of the sequencing error frequency in the DNA sequence databases. DNA Seq. **2:**343–346. (Erratum, **3:**337, 1993.)
146a. **Kumada, Y., E. Takano, K. Nagaoka, and C. J. Thompson.** 1990. *Streptomyces hygroscopicus* has two glutamine synthetase genes. J. Bacteriol. **172:**5343–5351.
147. **Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessières, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, and I. F. Connerton.** 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature **390:**249–256.
148. **Lake, J. A.** 1985. Evolving ribosome structure: domains in archaebacteria, eubacteria, eocytes and eukaryotes. Annu. Rev. Biochem. **54:**507–530.
149. **Lake, J. A.** 1988. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature **331:**184–186.
150. **Lake, J. A.** 1991. Tracing origins with molecular sequences: metazoan and eukaryotic beginnings. Trends Biochem. Sci. **16:**46–50.
151. **Lake, J. A.** 1991. The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. **8:**378–385.
152. **Lake, J. A.** 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91:**1455–1459.
153. **Lake, J. A., M. W. Clark, E. Henderson, S. P. Fay, M. Oakes, A. Scheinman, J. P. Thornber, and R. A. Mah.** 1985. Eubacteria, halobacteria, and the origin of photosynthesis: the photocytes. Proc. Natl. Acad. Sci. USA **82:**3716–3720.
154. **Lake, J. A., E. Henderson, M. W. Clark, and A. T. Matheson.** 1982. Mapping evolution with ribosome structure: intralineage constancy and interlineage variation. Proc. Natl. Acad. Sci. USA **79:**5948–5952.
155. **Lake, J. A., E. Henderson, M. Oakes, and M. W. Clark.** 1984. Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc. Natl. Acad. Sci. USA **81:**3786–3790.
156. **Lake, J. A., and M. C. Rivera.** 1994. Was the nucleus the first endosymbiont? Proc. Natl. Acad. Sci. USA **91:**2880–2881.
157. **Lancy, P., Jr., and R. G. E. Murray.** 1978. The envelope of Micrococcus radiodurans: isolation, purification and preliminary analysis of the wall

layers. Can. J. Microbiol. **24:**162–176.
158. **Langer, D., J. Hain, P. Thuriaux, and W. Zillig.** 1995. Transcription in archaea: similarity to that in eucarya. Proc. Natl. Acad. Sci. USA **92:**5768–5772.
159. **Margulis, L.** 1970. Origin of eukaryotic cells. Yale University Press, New Haven, Conn.
160. **Margulis, L.** 1992. Symbiosis theroy: cells as microbial communities, p. 149–172. *In* L. Margulis and L. Olendzenski (ed.), Environmental evolution: effects of the origin and evolution of life on planet earth. The MIT Press, Cambridge, Mass.
161. **Margulis, L.** 1992. Biodiversity: molecular biological domains, symbiosis and kingdom origins. Biosystems **27:**39–51.
162. **Margulis, L.** 1993. Symbiosis in cell evolution. W. H. Freeman & Co., New York, N.Y.
163. **Margulis, L.** 1996. Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. Proc. Natl. Acad. Sci. USA **93:**1071–1076.
164. **Margulis, L., and K. V. Schwartz.** 1988. Five kingdoms—an illustrated guide to the phyla of life on Earth. W. H. Freeman & Co., New York, N.Y.
165. **Martin, W., H. Brinkmann, C. Savonna, and R. Cerff.** 1993. Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc. Natl. Acad. Sci. USA **90:**8692–8696.
166. **Martin, W., and M. Muller.** 1998. The hydrogenosome hypothesis for the first eukaryote. Nature **392:**37–41.
167. **Mayr, E.** 1942. Systematics and the origin of species. Columbia University Press, New York, N.Y.
168. **Mayr, E.** 1968. The role of systematics in biology. Science **159:**595–599.
169. **Mayr, E.** 1990. A natural system of organisms. Nature **348:**491.
170. **Meyer, T. E., M. A. Cusanovich, and M. D. Kamen.** 1986. Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. Proc. Natl. Acad. Sci. USA **83:**217–220.
170a. **Morden, C. W., C. F. Delwiche, M. Kuhsel, and J. D. Palmer.** 1992. Gene phylogenies and the endosymbiotic origin of plastids. Biosystems **28:**75–90.
171. **Morell, V.** 1996. Life's last domain. Science **273:**1043–1045.
172. **Muller, M.** 1993. The hydrogenosome. J. Gen. Microbiol. **139:**2879–2889.
173. **Murray, R. G. E.** 1968. Microbial structure as an aid to microbial classification and taxonomy. Spisy Prirodoved. Fak. Univ. J. E. Purkyne Brne **43:**249–252.
174. **Murray, R. G. E.** 1984. Kingdom Procaryotae, p. 34–36. *In* N. R. Krieg and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 1. The Williams & Wilkins Co., Baltimore, Md.
175. **Murray, R. G. E.** 1984. The higher taxa, or, a place for everything . . .? p. 31–34. *In* N. R. Krieg and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 1. The Williams & Wilkins Co., Baltimore, Md.
176. **Murray, R. G. E.** 1986. Family II. *Deinococcaceae* Brooks and Murray 1981, 356^{VP}, p. 1035–1043. *In* P. H. A. Sneath, N. S. Mair, M. E. Sharpe, and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 2. The Williams & Wilkins Co., Baltimore, Md.
177. **Murray, R. G. E., D. J. Brenner, R. R. Colwell, P. De Vos, M. Goodfellow, P. A. D. Grimont, N. Pfennig, E. Stackebrandt, and G. A. Zavarzin.** 1990. Report of the Ad Hoc Committee on Approaches to Taxonomy within the Proteobacteria. Int. J. Syst. Bacteriol. **40:**213–215.
178. **Nei, M.** 1991. Relative efficiencies of different tree-making methods for molecular data, p. 90–128. *In* M. M. Miyamoto and J. Cracraft (ed.), Phylogenetic analysis of DNA sequences. Oxford University Press, New York, N.Y.
179. **Neu, H. C.** 1992. The crisis in antibiotic resistance. Science **257:**1064–1072.
180. **Nicholson, R. C., D. B. Williams, and L. A. Moran.** 1990. An essential member of the *HSP70* gene family of *Saccharomyces cerevisiae* is homologous to immunoglobulin heavy chain binding protein. Proc. Natl. Acad. Sci. USA **87:**1159–1163.
180a. **Nikaido, H.** 1994. Prevention of drug access to bacterial targets: permeability barriers and active efflux. Science **264:**382–387.
180b. **Nikaido, H., S.-H. Kim, and E. Y. Rosenberg.** 1993. Physical organization of lipids in the cell wall of *Mycobacterium chelonae*. Mol. Microbiol. **8:**1025–1030.
181. **Olsen, G. J., and C. R. Woese.** 1993. Ribosomal RNA: a key to phylogeny. FASEB J. **7:**113–123.
182. **Olsen, G. J., and C. R. Woese.** 1996. Lessons from an Archaeal genome: what are we learning from Methanococcus jannaschii? Trends Genet. **12:**377–379.
183. **Olsen, G. J., and C. R. Woese.** 1997. Archaeal genomics: an overview. Cell **89:**991–994.
184. **Olsen, G. J., C. R. Woese, and R. Overbeek.** 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. **176:**1–6.
185. **Opperman, T., and J. P. Richardson.** 1994. Phylogenetic analysis of sequences from diverse bacteria with homology to the *Escherichia coli rho* gene. J. Bacteriol. **176:**5033–5043.
186. **Pace, N. R.** 1991. Origin of life—facing up to the physical setting. Cell **65:**531–533.
187. **Pace, N. R.** 1997. A molecular view of microbial diversity and the biosphere. Science **276:**734–740.
188. **Pace, N. R., G. J. Olsen, and C. R. Woese.** 1986. Ribosomal RNA phylogeny and the primary lines of evolutionary descent. Cell **45:**325–326.
189. **Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen.** 1986. The analysis of natural microbial populations by ribosomal RNA sequences, p. 1–55. *In* K. C. Marshall (ed.), Advances in microbial ecology. Plenum Press, New York, N.Y.
190. **Pelham, H. R. B.** 1989. Heat shock and the sorting of luminal ER proteins. EMBO J. **8:**3171–3176.
191. **Pennisi, E.** 1998. Genome data shake tree of life. Science **280:**672–674.
192. **Perry, J. J., and J. T. Staley.** 1996. Microbiology: dynamics and diversity. Saunders College Publishing, Philadelphia, Pa.
193. **Pratt, W. B.** 1993. The role of heat shock proteins in regulating the function, folding and trafficking of the glucocorticoid receptor. J. Biol. Chem. **268:**21455–21458.
194. **Prevot, A. R.** 1940. Manuel de classification et de determination des bacteres anaerobies. Masson et Cie, Paris, France.
195. **Pringsheim, E. G.** 1949. The relationship between bacteria and Myxophycae. Bacteriol. Rev. **13:**47–98.
196. **Puhler, G., H. Leffers, F. Gropp, P. Palm, H. P. Klenk, F. Lottspeich, R. A. Garrett, and W. Zillig.** 1989. Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. Proc. Natl. Acad. Sci. USA **86:**4569–4573.
197. **Reeve, J. N., K. Sandman, and C. J. Daniels.** 1997. Archaeal histones, nucleosomes, and transcription initiation. Cell **89:**999–1002.
197a. **Ribeiro, S., and G. B. Golding.** 1998. The mosaic nature of the eukaryotic nucleus. Mol. Biol. Evol. **15:**779–788.
198. **Rivera, M. C., and J. A. Lake.** 1992. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science **257:**74–76.
199. **Roger, A. J., C. G. Clark, and W. F. Doolittle.** 1996. A possible mitochondrial gene in the early-branching amitochondriate protist *Trichomonas vaginalis*. Proc. Natl. Acad. Sci. USA **93:**14618–14622.
200. **Roger, A. J., S. G. Svärd, J. Tovar, C. G. Clark, M. W. Smith, F. D. Gillin, and M. L. Sogin.** 1998. A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. Proc. Natl. Acad. Sci. USA **95:**229–234.
201. **Rosenthal, B., Z. Mai, D. Caplivski, S. Ghosh, H. De La Vega, T. Graf, and J. Samuelson.** 1997. Evidence for the bacterial origin of genes encoding fermentation enzymes of the amitochondriate protozoan parasite *Entamoeba histolytica*. J. Bacteriol. **179:**3736–3745.
202. **Rothschild, L. J., M. A. Ragan, A. W. Coleman, P. Heywood, and S. A. Gerbi.** 1986. Are rRNA sequence comparisons the Rosetta stone of phylogenetics? Cell **47:**640.
203. **Rowlands, T., P. Baumann, and S. P. Jackson.** 1994. The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria. Science **264:**1326–1329.
204. **Russell, A. D., and I. Chopra.** 1990. Understanding antibacterial action and resistance. Ellis Horwood, New York, N.Y.
205. **Saccone, C., C. Gissi, C. Lanave, and G. Pesole.** 1995. Molecular classification of living organisms. J. Mol. Evol. **40:**273–279.
206. **Saier, M. H., Jr.** 1979. The role of the cell surface in regulating the internal environment, p. 167–227. *In* J. R. Sokatch and L. N. Ornston (ed.), The bacteria, vol. VII. Academic Press, Inc., New York, N.Y.
207. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
208. **Schopf, J. W.** 1978. The evolution of the earliest cells. Sci. Am. **239:**110–120.
209. **Schopf, J. W.** 1994. Disparate rates, differing fates: tempo and mode of evolution changed from the Precambrian to the Phanerozoic. Proc. Natl. Acad. Sci. USA **91:**6735–6742.
210. **Schubert, I.** 1988. Eukaryotic nuclei of endosymbiontic origin? Naturwissenschaften **75:**89–91.
211. **Schwartz, R. M., and M. O. Dayhoff.** 1978. Origin of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science **199:**395–403.
212. **Searcy, D. G.** 1982. *Thermoplasma*: a primordial cell from a refuse pile. Trends Biochem. Sci. **7:**183–185.
213. **Shimmin, L. C., C. Ramirez, A. T. Matheson, and P. P. Dennis.** 1989. Sequence alignment and evolutionary comparison of the L10 equivalent and L12 equivalent ribosomal proteins from archaebacteria, eubacteria, and eucaryotes. J. Mol. Evol. **29:**448–462.
214. **Singh, B., B. J. Soltys, Z. C. Wu, H. V. Patel, K. B. Freeman, and R. S. Gupta.** 1997. Cloning and some novel characteristics of mitochondrial Hsp70 from Chinese hamster cells. Exp. Cell Res. **234:**205–216.
215. **Smith, D. R., L. A. Doucette-Stamm, C. Deloughery, H. M. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Y. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gibson, N. Jiwani, A. Caruso, and D. Bush.** 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* DeltaH: functional analysis and comparative genomics. J. Bacteriol. **179:**7135–7155.
216. **Smith, M. W., D. F. Feng, and R. F. Doolittle.** 1992. Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem. Sci. **17:**489–493.

217. **Sogin, M. L.** 1991. Early evolution and the origin of eukaryotes. Curr. Opin. Genet. Dev. **1:**457–463.
218. **Sogin, M. L.** 1997. History assignment: when was the mitochondrion found. Curr. Opin. Genet. Dev. **7:**792–799.
219. **Sogin, M. L., J. H. Gunderson, H. J. Elwood, R. A. Alonso, and D. A. Peattie.** 1989. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. Science **243:**75–77.
220. **Sokatch, J. R.** 1979. Roles of appendages and surface layers in adaptation of bacteria to their environment, p. 229–289. *In* L. N. Ornston and J. R. Sokatch (ed.), The bacteria, vol. VII. Academic Press, Inc., New York, N.Y.
221. **Soltys, B. J., M. Falah, and R. S. Gupta.** 1996. Identification of endoplasmic reticulum in the primitive eukaryote Giardia lamblia using cryoelectron microscopy and antibody to Bip. J. Cell Sci. **109:**1909–1917.
222. **Soltys, B. J., and R. S. Gupta.** 1994. Presence and cellular distribution of a 60-kDa protein related to mitochondrial hsp60 in Giardia lamblia. J. Parasitol. **80:**580–590.
223. **Spratt, B. G.** 1994. Resistance to antibiotics mediated by target alterations. Science **264:**388–393.
224. **Stackebrandt, E.** 1992. Unifying phylogeny and phenotypic diversity, p. 19–47. *In* A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K. H. Schleifer (ed.), The prokaryotes, 2nd ed. Springer-Verlag, New York, N.Y.
225. **Stackebrandt, E., R. G. E. Murray, and H. G. Trüper.** 1988. *Proteobacteria* classic nov., a name for the phylogenetic taxon that includes the "purple bacteria and their relatives." Int. J. Syst. Bacteriol. **38:**321–325.
226. **Stackebrandt, E., and C. R. Woese.** 1984. The phylogeny of prokaryotes. Microbiol. Sci. **1:**117–122.
227. **Stanier, R. Y.** 1941. The main outlines of bacterial classification. J. Bacteriol. **42:**437–466.
228. **Stanier, R. Y., E. A. Adelberg, and J. L. Ingraham.** 1976. The microbial world. Prentice-Hall, Inc., Englewood Cliffs, N.J.
229. **Stanier, R. Y., J. L. Ingraham, M. L. Wheelis, and P. R. Painter.** 1987. General microbiology. Macmillan Education Ltd., London, England.
230. **Stanier, R. Y., and C. B. van Niel.** 1962. The concept of a bacterium. Arch. Mikrobiol. **42:**17–35.
231. **Steel, M. A., P. J. Lockhart, and D. Penny.** 1993. Confidence in evolutionary trees from biological sequence data. Nature **364:**440–442.
232. **Stetter, K. O.** 1995. Microbial life in hyperthermal environments. ASM News **61:**285–290.
233. **Stewart, C.-B.** 1993. The powers and pitfalls of parsimony. Nature **361:**603–607.
234. **Stiller, J. W., and B. D. Hall.** 1997. The origin of red algae: implications for plastid evolution. Proc. Natl. Acad. Sci. USA **94:**4520–4525.
235. **Suzuki, K., M. Goodfellow, and A. G. O'Donnell.** 1993. Cell envelopes and classification, p. 195–250. *In* M. Goodfellow and A. G. O'Donnell (ed.), Handbook of new bacterial systematics. Academic Press, Inc., New York, N.Y.
236. **Swofford, D. L., and G. L. Olsen.** 1990. Phylogeny reconstruction, p. 411–501. *In* D. Hillis and C. Moritz (ed.), Molecular systematics. Sinauer Associates, Inc., Sunderland, Mass.

236a. **Syvanen, M.** 1994. Horizontal gene transfer: evidence and possible consequences. Annu. Rev. Genet. **28:**237–261.

237. **Szathmary, E., and J. M. Smith.** 1995. The major evolutionary transitions. Nature **374:**227–232.
238. **Tateno, Y., N. Takezei, and M. Nei.** 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **12:**261–277.
239. **Tiboni, O., P. Cammarano, and A. M. Sanangelantoni.** 1993. Cloning and sequencing of the gene coding glutamine synthetase I from the archaeum *Pyrococcus woesi*: anomalous phylogenies inferred from analysis of archael and bacterial glutamine synthetase I sequence. J. Bacteriol. **175:**2961–2969.
240. **Tiboni, O., R. Cantoni, R. Creti, P. Cammarano, and A. M. Sanangelantoni.** 1991. Phylogenetic depth of Thermotoga maritima inferred from analysis of the fus gene: amino acid sequence of elongation factor G and organization of the Thermotoga str operon. J. Mol. Evol. **33:**142–151.
241. **Tipper, D. J., and A. Wright.** 1979. The structure and biosynthesis of bacterial cell walls, p. 291–415. *In* J. R. Sokatch and L. N. Ornston (ed.), The bacteria, vol. VII. Academic Press, Inc., New York, N.Y.
242. **Tomb, J. F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E. F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzegerald, N. Lee, M. D. Adams, J. C. Venter, et al.** 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature **388:**539–547.
243. **Trent, J. D., E. Nimmesgern, J. S. Wall, F. U. Hartl, and A. L. Horwich.** 1991. A molecular chaperone from a thermophilic archaebacterium is related to the eukaryotic protein t-complex polypeptide 1. Nature **354:**490–493.
244. **Trüper, H. G., and K. H. Schleifer.** 1992. Prokaryote characterization and identification, p. 126–148. *In* A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K. H. Schleifer (ed.), The prokaryotes, 2nd ed. Springer-Verlag, New York, N.Y.
245. **van Niel, C. B.** 1946. The classification and natural relationships of bacteria. Cold Spring Harbor Symp. Quant. Biol. **11:**285–301.
246. **Viale, A. M., A. K. Arakaki, F. C. Soncini, and R. G. Ferreyra.** 1994. Evolutionary relationships among eubacterial groups as inferred from GroEL (chaperonin) sequence comparisons. Int. J. Syst. Bacteriol. **44:**527–533.

246a. **Viale, A. M., and A. K. Arakaki.** 1994. The chaperone connection to the origins of the eukaryotic organelles. FEBS Lett. **341:**146–151.

247. **Wetmur, J. G., D. M. Wong, B. Ortiz, J. Tong, F. Reichert, and D. H. Gelfand.** 1994. Cloning, sequencing, and expression of RecA proteins from three distantly related thermophilic eubacteria. J. Biol. Chem. **269:**25928–25935.
248. **Whittaker, R. H., and L. Margulis.** 1978. Protist classification and the kingdoms of organisms. Biosystems **10:**3–18.
249. **Winefield, C. S., K. J. Farnden, P. H. Reynolds, and C. J. Marshall.** 1995. Evolutionary analysis of aspartate aminotransferases. J. Mol. Evol. **40:**455–463.

249a. **Woese, C. R.** 1981. Archaebacteria. Sci. Am. **244:**98–122.

250. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221–271.
251. **Woese, C. R.** 1991. The use of ribosomal RNA in reconstructing evolutionary relationships among bacteria, p. 1–24. *In* R. K. Selander, A. G. Clark, and T. S. Whittmay (ed.), Evolution at the molecular level. Sinauer Associates, Inc., Sunderland, Mass.
252. **Woese, C. R.** 1992. Prokaryote systematics: the evolution of a science, p. 3–18. *In* A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K. H. Schleifer (ed.), The prokaryotes, 2nd ed. Springer-Verlag, New York, N.Y.
253. **Woese, C. R.** 1993. The Archaea: their history and significance, p. vii–xxix. *In* M. Kates, D. J. Kushner, and A. T. Matheson (ed.), The biochemistry of Archaea (Archaebacteria). Elsevier Science Publishers B.V., New York, N.Y.
254. **Woese, C. R.** 1994. There must be a prokaryote somewhere: microbiology's search for itself. Microbiol. Rev. **58:**1–9.
255. **Woese, C. R.** 1998. The universal ancestor. Proc. Natl. Acad. Sci. USA **95:**6854–6859.
256. **Woese, C. R., and G. E. Fox.** 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA **74:**5088–5090.
257. **Woese, C. R., R. Gutell, R. Gupta, and H. F. Noller.** 1983. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. Microbiol. Rev. **47:**621–669.
258. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:**4576–4579.
259. **Woese, C. R., L. J. Magrum, R. Gupta, R. B. Siegel, D. A. Stahl, J. Kop, N. Crawford, J. Brosius, R. Gutell, J. J. Hogan, and H. F. Noller.** 1980. Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. Nucleic Acids Res. **8:**2275–2293.
260. **Wright, A., and D. J. Tipper.** 1979. The outer membrane of gram-negative bacteria, p. 427–485. *In* J. R. Sokatch and L. N. Ornston (ed.), The bacteria, vol. VII. Academic Press, Inc., New York, N.Y.
261. **Yang, D., Y. Oyaizu, H. Oyaizu, G. J. Olsen, and C. R. Woese.** 1985. Mitochondrial origins. Proc. Natl. Acad. Sci. USA **82:**4443–4447.
262. **Zillig, W.** 1991. Comparative biochemistry of Archaea and Bacteria. Curr. Opin. Genet. Dev. **1:**544–551.
263. **Zillig, W., R. Schnabel, and K. O. Stetter.** 1985. Archaebacteria and the origin of the eukaryotic cytoplasm. Curr. Top. Microbiol. Immunol. **114:**1–18.
264. **Zuckerkandl, E., and L. Pauling.** 1965. Molecules as documents of evolutionary history. J. Theor. Biol. **8:**357–366.