Theoretical Population Biology **61**, 423–434 (2002)
doi:10.1006/tpbi.2002.1589

TPB

Critical Issues in Bacterial Phylogeny

Radhey S. Gupta[1] and Emma Griffiths
Department of Biochemistry, McMaster University, Hamilton, Ontario, Canada L8N 3Z5
Received April 10, 2002

To understand bacterial phylogeny, it is essential that the following two critical issues be resolved: (i) development of well-defined (molecular) criteria for identifying the main groups within *Bacteria*, and (ii) to understand how the different main groups are related to each other and how they branched off from a common ancestor. These issues are not resolved at present. We have recently described a new approach, based on shared conserved inserts and deletions (indels or signature sequences) found in various proteins, that provides a reliable means for understanding these issues. A large number of conserved indels that are shared by different groups of bacteria have been identified. Using these indels, and based simply on their presence or absence, all of the main groups within *Bacteria* can be defined in clear molecular terms and new species could be assigned to them with minimal ambiguity. The analysis of these indels also permits one to logically deduce that the various main bacterial groups have branched off from a common ancestor in the following order: Low G+C Gram-positive ⇒ High G+C Gram-positive ⇒ *Clostridium–Fusobacteria–Thermotoga* ⇒ *Deinococcus–Thermus*–Green nonsulfur bacteria ⇒ Cyanobacteria ⇒ *Spirochetes* ⇒ *Chlamydia–Cytophaga–Bacteroides*–Green sulfur bacteria ⇒ *Aquifex* ⇒ Proteobacteria 1 (ε and δ) ⇒ Proteobacteria-2. (α) ⇒ Proteobacteria-3 (β) and ⇒ Proteobacteria-4 (γ). The validity of this approach was tested using sequence data from bacterial genomes. By making use of 18 conserved indels, species from all 60 completed bacterial genomes were assigned to different groups. The observed distribution of these indels in different species was then compared with that predicted by the model. Of the 936 observations concerning the placement of these indels in various species, all except one were in accordance with the model. The placement of bacteria into different groups using this approach also showed excellent correlation with the 16S rRNA phylogenies with nearly all of the species assigned to the same groups by both methods. These results provide strong evidence that the genes containing these indels have not been affected by factors such as lateral gene transfers. However, such events are readily detected by this means and some examples are provided. The approach described here thus provides a reliable and internally consistent means for understanding various critical and long outstanding issues in bacterial phylogeny. © 2002 Elsevier Science (USA)

INTRODUCTION

An understanding of the evolutionary relationships among prokaryotic organisms constitutes a long cherished goal of microbiology and biological sciences (Kluyver and van Niel, 1936; Stanier, 1941; Stanier and van Niel, 1962). In view of their ancestral nature, an understanding of the evolutionary relationships among them is critical for understanding numerous fundamental questions such as the nature and origin of the first cell, origin of metabolisms, origin of photosynthesis and of the information transfer processes, and also the origin of the ancestral eukaryotic cell. There have been numerous attempts made in the past to group and classify prokaryotes based on resemblances in their morphological, biochemical and physiological characteristics (Orla-Jensen, 1909; Buchanan, 1925; Kluyver and van

[1]To whom correspondence should be addressed. E-mail: gupta@mcmaster.ca.

0040-5809/02 $35.00
© 2002 Elsevier Science (USA)
All rights reserved.

Niel, 1936; Stanier, 1941; Stanier and van Niel, 1962; Murray, 1986a). For example, the eighth edition of *Bergey's Manual of Determinative Bacteriology* published in 1974 (Buchanan and Gibbons, 1974), empirically divided prokaryotes into two major divisions: Division I consisting of Cyanobacteria which are capable of carrying out oxygenic photosynthesis and Division II consisting of all other bacteria. The latter division was further divided into 19 groups, such as Gliding bacteria, Sheathed bacteria, Spirochetes, Gram-negative aerobic rods and cocci, Gram-negative anaerobic bacteria, Methane-producing bacteria, Gram-positive cocci, Endospore-forming rods and cocci, Actinomycetes and related organisms, Rickettsias, Mycoplasma, etc. Although the division of prokaryotes into these groups was helpful for deterministic purposes, it was not clear whether the species in these groups were indeed evolutionarily related (Kluyver and van Niel, 1936; Stanier, 1941; Stanier and van Niel, 1962). The questions as to how these groups were related to each other or evolved from a common ancestor were not even seriously considered at the time.

A major change in this regard came about with the recognition that the linear sequences of macromolecules in different species contain an enormous reservoir of both qualitative and quantitative characters deriving directly from the common ancestor (Zuckerkandl and Pauling, 1965). Based on alignment of sequences for homologous genes/proteins, the number and nature of sequence changes can be determined between different species. This formed the basis for deducing the genealogical relationships among species based on molecular sequences. The earliest detailed application of this approach by Carl Woese and colleagues using 16S rRNA sequences led to the suggestion that the prokaryotic organisms are of two different kinds, archaebacteria (or *Archaea*) and eubacteria (or *Bacteria*), which are distinct from each other and originated independently from a universal ancestor (Fox *et al*., 1980; Woese, 1987). Although *Archaea* are distinct from *Bacteria* with regard to a large number of characteristics, the phylogenetic relationship between these groups is controversial. The controversy surrounding this issue has been discussed in detail previously (Gupta, 1998 a–c, 2000a). and it will not be dealt with here. The main objective of this article is to discuss issues that are critical to understanding the evolutionary relationships among *Bacteria*.

BACTERIAL PHYLOGENY: CURRENT UNDERSTANDING AND KEY UNRESOLVED ISSUES

Bacteria comprise the vast majority of the known prokaryotes and hence an understanding of the relationship among them constitutes a major part of the prokaryotic phylogeny. The two central issues in terms of understanding bacterial phylogeny are: (i) Description of well-defined molecular criteria for identifying the main groups or divisions among *Bacteria*. These criteria should enable one to place any given species into one of the defined groups in an unambiguous manner. A related issue is to describe criteria for identifying the major subdivisions within a given group. (ii) To develop means to understand how the different main groups are related to each other and how they branched off from a common ancestor. The understanding of these issues is critical for placing the bacterial phylogeny on a firm footing. We will first examine our present understanding (or lack thereof) of these issues and subsequently describe new approaches by which these could be resolved.

Our current understanding of the evolutionary relationships among *Bacteria* is largely based on 16S rRNA sequences (Woese, 1987; Balows *et al*., 1992; Olsen *et al*., 1994; Ludwig and Klenk, 2001). Based on oligonucleotide signatures and branching patterns in the 16S rRNA trees, 10–12 main groups or divisions within bacteria were initially proposed (Fox *et al*., 1980; Woese *et al*., 1985; Woese, 1987). These included, thermotoga, green nonsulfur bacteria, *Deinococci* and relatives, cyanobacteria, low G + C Gram-positive (*Firmicutes*), high G + C Gram-positive (*Actinobacteria*), spirochetes, green sulfur bacteria, cytophaga, chlamydiae, planctomycetes and proteobacteria. The basis of suggesting these divisions was their distinct branching in the rRNA trees. At the time when these groups were proposed, the sequence database was quite limited, and these groups could be clearly distinguished from each other based on long 'naked' internal branches that separated them. However, during the past 10–15 years, the explosive increase in available sequence data (Maidak *et al*., 2001) has effectively 'filled-in' these once naked branches, making the distinction between these groups on the basis of their branching in the rRNA trees increasingly difficult and imprecise (Ludwig and Schleifer, 1999; Ludwig and Klenk, 2001).

In recent years, many additional groups or divisions within *Bacteria* have been suggested (viz., *Aquificales*, *Desulfurobacterium*, *Thermomicrobia*, *Chrysiogenetes*,

Deferribacters Dictyoglomus, *Fusobacteria*, *Holophaga*, *Fibrobacter*, *Nitrospira*, *Flexistipes*, *Verrucomicrobium*) (Ludwig and Klenk, 2001). Some of these groups consist of only one or a few species (e.g., *Thermomicrobia*, *Chrysiogenetes*, *Fibrobacteres*, *Deferribacters*), and they were earlier parts of other divisions. In contrast, a number of other groups such as proteobacteria, cyanobacteria, *Actinobacteria* (high G + C Gram-positive) and *Firmicutes* (low G + C Gram-positive) are very large and are composed of several hundred to thousands of species accounting for more than 90–95% of all known bacteria (Holt, 1984). In the absence of clear and objective criteria for defining the major divisions, and the problems associated with branching of species in phylogenetic trees, it is unclear how many of these recently described groups actually constitute major *new* divisions within *Bacteria*. To place the bacterial phylogeny on a firmer basis, it is thus essential to develop new more objective criteria for defining the main groups or divisions within *Bacteria*, which give stable relationships.

In this context, it is also important to develop objective criteria for describing the major subdivisions within a given group. Currently, based on 16S rRNA trees, the proteobacterial group has been divided into five subdivisions, namely $\alpha, \beta, \gamma, \delta$ and ε (Woese, 1987; Stackebrandt *et al*., 1988). Some of these subdivisions, viz., α, β and γ, consist of several hundreds to thousands of species and they are much larger than most other known main divisions within *Bacteria*. The member species of these subdivisions can be clearly distinguished from each other, and from other bacterial divisions, both in phylogenetic trees and by means of a large number of distinctive signature sequences in proteins (Olsen *et al*., 1994; Viale *et al*., 1994; Karlin *et al*., 1995; Eisen, 1995; Gupta, 2000b). In the absence of any clearly stated criteria for defining divisions or subdivisions, it is unclear why these major bacterial groups have been assigned a subdivision status, whereas many poorly characterized groups consisting of a single species are recognized as separate divisions.

The second critical issue for understanding bacterial phylogeny is to determine how the different main groups or divisions within *Bacteria* are related to each other and how they branched from a common ancestor. Such relationships are not resolved in phylogenetic trees based on rRNA or different proteins, and this limitation of phylogenetic trees is widely accepted (Woese, 1991; Olsen and Woese, 1993; Brendel *et al*., 1997; Gupta, 1998a; Doolittle, 1999; Ludwig and Schleifer, 1999; Ludwig and Klenk, 2001). The inability of the 16S rRNA or various proteins trees to resolve these relationships has generally been assumed to mean that this important problem is basically insolvable. This has led to an increasing acceptance of the notion that all (or most) main groups within *Bacteria* branched off from a common ancestor at a similar time (Doolittle, 1999; Ludwig and Schleifer, 1999; Ludwig and Klenk, 2001). However, how different groups of bacteria are related to each other and in what order they have evolved from a common ancestor, forms the crux of bacterial phylogeny and in the absence of any knowledge regarding these aspects, our understanding of bacterial phylogeny remains largely incomplete and superficial.

The inability of phylogenetic trees to resolve these key issues can be largely attributed to the dependence of the derived inferences upon many different variables and assumptions. These include, reliability of the sequence data and alignment, regions of the sequences that are retained or excluded in phylogenetic analysis, number and range of species examined, order of addition of species in generating sequence alignment, differences in the evolutionary rates and base compositions of the species, phylogenetic methods employed, etc. (Woese, 1987; Lake, 1991; Gupta, 1998a; Ludwig and Klenk, 2001.) There are no standard criteria for generating optimal sequence alignment or for constructing phylogenetic trees. Often a change in only one of the above parameters can have a marked effect on the branching patterns of species in phylogenetic trees. Although the effect of these variables on the relationships among closely related species is generally small, it can greatly alter the relative branching orders of the major lineages (Ludwig and Klenk, 2001). The difficulty in controlling and assessing the effects of these variables on branching patterns in phylogenetic trees is widely recognized as exemplified by this statement from Woese (1991), ''When it comes to defining the branching order among major groups, or to defining many of the lesser taxa within them, one is at the mercy of powerful and sophisticated tree construction algorithms and can consequently fall prey to their vagaries''. Hence, to understand the evolutionary relationships among *Bacteria*, it is essential that new sequence-based criteria be developed that are minimally affected by these variables, but which are capable of resolving distant evolutionary relationships. A new approach that is showing great promise in these regards involves the use of shared conserved inserts and deletions (also referred to as signature sequences) found in various proteins (Gupta, 1998a, 2000b, 2001). The rationale of this approach and how it has proven useful in understanding bacterial phylogeny is discussed below.

THE UTILITY OF CONSERVED INDELS AS PHYLOGENETIC MARKERS TO DEFINE THE MAIN BACTERIAL TAXA AND THEIR BRANCHING ORDERS

When a conserved indel (i.e., either insert or deletion) of defined length and sequence is found at the same position in a given protein (or gene) in all members from one or more groups of bacteria, then the simplest and most parsimonious explanation for this observation is that the indel was introduced only once in a common ancestor of these species. Based on the presence or absence of shared indels, different species can be divided into distinct groups, those containing or lacking the indel (Gupta, 1998a, 2000b). Well-defined indels in protein sequences also serve as useful milestones for evolutionary events, since it is expected that all species emerging from an ancestral cell in which the indel was first introduced will contain the signature, whereas all species that either existed prior to this event or which did not evolve from this ancestor will lack the indel. To interpret whether a given indel is the result of an insertion or deletion event, and to understand the evolutionary significance of a given signature concerning the ancestral or derived nature of the groups, a reference point is essential (Gupta, 1998a,b). The root of the prokaryotic tree provides a very useful reference point for such purposes. Based on duplicated gene sequences for elongation factor-1 and-2 (Iwabe *et al*., 1989), sequence characteristics of the Hsp70 proteins (Gupta and Singh, 1992) and the cell structures of prokaryotic organisms, it has previously been suggested that the root of the prokaryotic tree lies between archaebacteria and Gram-positive bacteria (Gupta, 1998a). Within prokaryotes, only these two groups are bounded by a single unit lipid membrane. In contrast, all other bacteria (i.e., Gram-negative bacteria) contain both an inner as well as an outer cell membrane. (The Gram-positive bacteria are defined in our work as those surrounded by a single membrane and not solely by their Gram-staining characteristics (Gupta, 1998b).) Since membrane enclosure was likely a key event in the formation of the ancestral prokaryotic cell (Morowitz, 1992), it is reasonable to expect that the earliest prokaryotic cell was bounded by a simpler cell enclosure as found in the Gram-positive bacteria and archaebacteria, rather than two different membranes separated by an intervening compartment. The ancestral nature of these groups of prokaryotes is also supported by other lines of evidence discussed elsewhere (Gupta, 1998a). Based on this rooting, it is now possible to use the signature sequences found in various proteins to logically deduce the branching order of different groups of bacteria.

Our work has led to the identification of a large number of conserved indels in different proteins which provide valuable markers for such studies (Gupta, 1998a, 2000b; Griffiths and Gupta, 2002). Some of these signatures which have proven most useful for determining the relative branching orders of different groups of bacteria are shown in Fig. 1. The detailed descriptions of most of these signatures can be found in our published work (Gupta, 1998a, 2000b, 2001) and also on the faculty webpage of RSG (http://www.science.mcmaster.ca/biochem/). The analyses of these signatures indicate that they have been introduced at the indicated specific stages of bacterial evolution as shown in Fig. 1. Based upon the presence or absence of these indels, most of the major groups within *Bacteria* can be clearly distinguished and their relative branching order from a common ancestor can be established (Gupta, 2001).

An example of a newly identified signature sequence in the transcription termination factor rho is shown in Fig. 2. The rho protein is present in all main groups of bacteria, except cyanobacteria, where this gene has presumably been lost. The signature in this protein consists of a 2 a.a. insert in a conserved region which is commonly shared by different groups of proteobacteria, *Aquifex, Cytophaga–Flavobacteria*–chlamydiae and spirochetes. However, this insert is not found in the *Deinococcus/Thermus* group, green nonsulfur bacteria and in different groups of Gram-positive bacteria. In our scheme, this signature is postulated to have been introduced in a common ancestor of proteobacteria, *Aquifex, Cytophaga–Flavobacteria*–chlamydiae and spirochetes groups of species after the branching of the other groups. The presence of this insert in *T. maritima* constitutes an exception in this case, which could be a consequence of either lateral gene transfer (LGT) or independent occurrence of this indel in this particular species.

The evolutionary model based on indels (Fig. 1) was developed based on limited sequence data (Gupta, 1998a). However, in the past 3–4 years, sequence data for a large number of bacterial genomes have become available. This provides a very powerful and objective means to test the reliability of the proposed model. At the present time, genome sequences for 60 bacterial species are known (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html). The presence or absence of all 18 signatures shown in Fig. 1 in these species was examined by means of BLAST analysis and sequence alignments. Based on the presence or absence of these

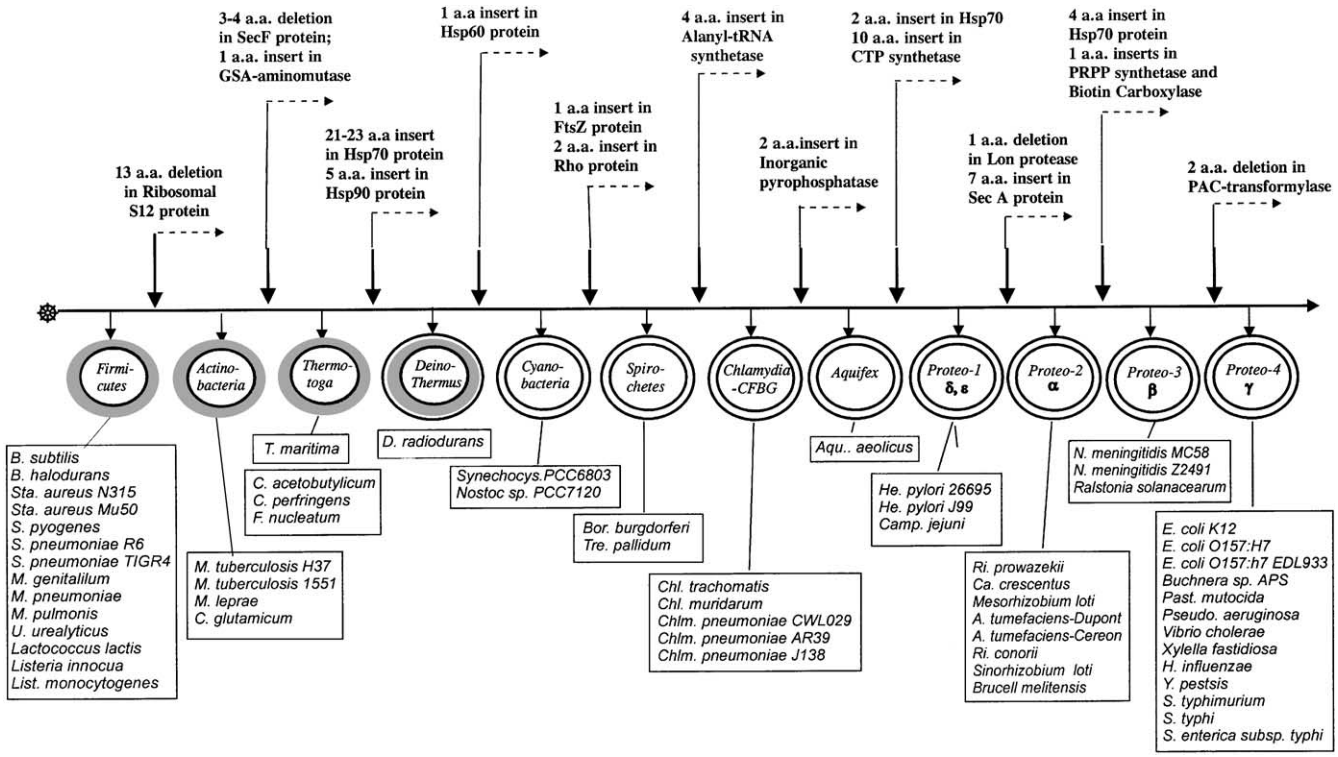

FIG. 1. Phylogenetic placement and the relative branching order of bacterial species from completed genomes based on signature sequences in different proteins. The thick arrows above the line show the specific stages where the indicated indels are postulated to have been introduced. The model predicts that all bacterial groups to the right of these arrows should contain the indicated signatures whereas all groups on the left should lack them. The sequences from completed bacterial genomes strongly conform to the expected patterns with only a single exception observed in 936 observations concerning the placement (i.e., presence or absence) of indels.

signatures, the bacterial species were assigned to various identified groups. The reliability of the model could be tested by comparing the observed distribution of these indels in different species with that predicted by the model (Fig. 1). Once an indel has been introduced in an ancestral lineage, various groups emerging after that point should all contain the indel, whereas all species that existed prior to the introduction of the indel, should lack the signature. The model thus makes very specific predictions as to which of these signatures should be present or absent in different species (Gupta, 1998a, 2001). In contrast, if this model was unreliable and if the various identified indels were introduced either independently or if the genes containing them have undergone lateral gene transfer from one species to another, then the presence or absence of these markers in different species will not follow the predicted pattern. In such a case, different groups of species or even individual species from different groups will either contain or lack the indels, yielding contradictory results.

Results of these analyses for various indels are presented in Table I. The table lists the number of completed genomes where these proteins are found, and also the number of species that are predicted to contain each of the identified signatures. For example, for the signature sequence in Ala-tRNA synthetase, a protein found in all sequenced bacterial genomes, the model predicts that 33 species should contain this indel whereas the remainder 27 should not have it. This is exactly what is found. The last few columns in this table summarize the actual results obtained for different indels and the number of exceptions or contradictions that are observed. Some of these genes/proteins are not found in all species, hence the total number is not the same in all cases but corresponds to the number of species where these genes are present in the genomes. The results of these studies are remarkably clear (Table I), as the presence or absence of these signatures in different genomes followed exactly the pattern as predicted by the model. In a total of 936 observations involving the placement of indels in different species, only one

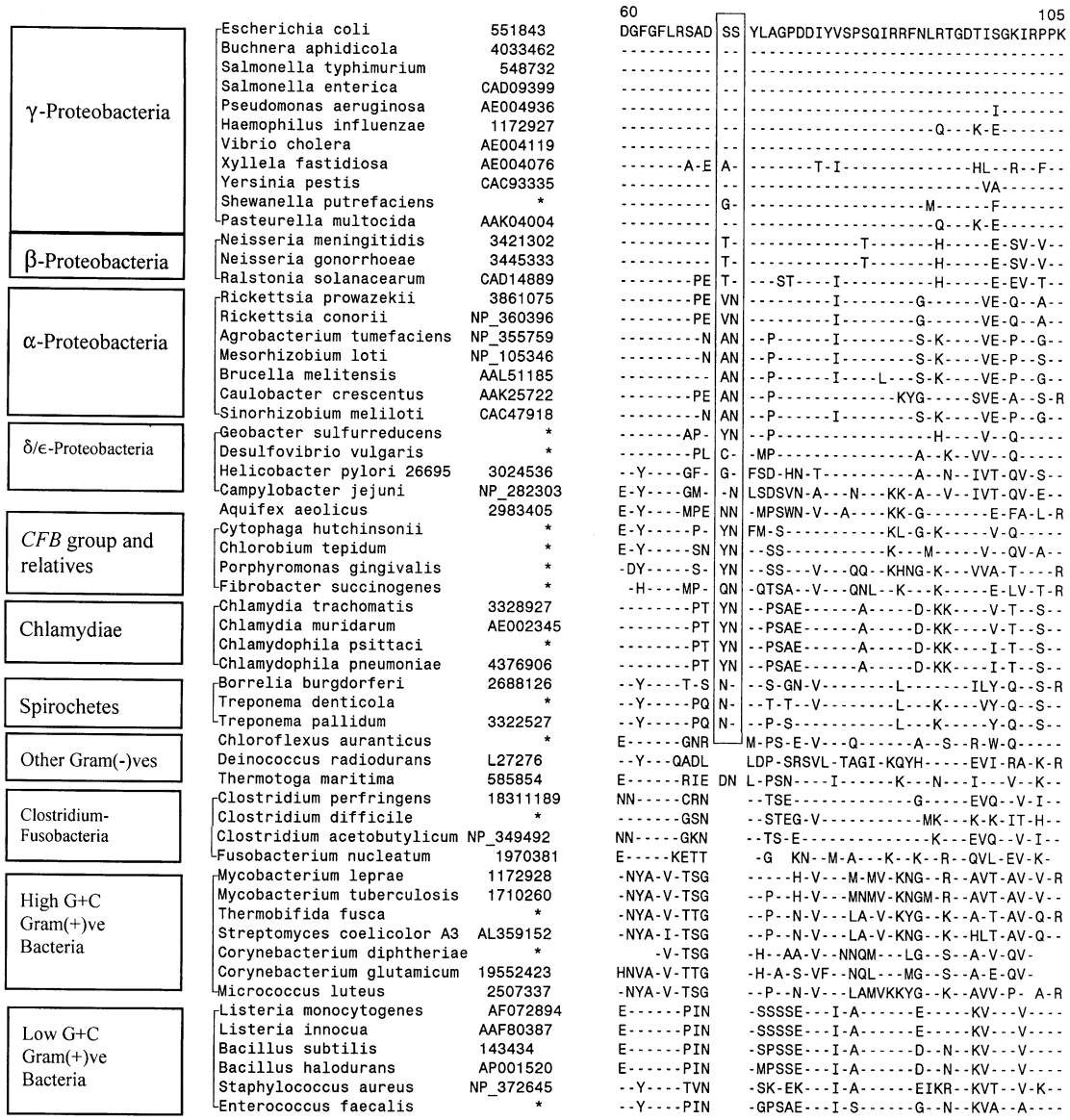

FIG. 2. A signature sequence in transcription factor rho that is commonly shared by various proteobacteria, *Aquifex* and chlamydiae, *CFB* group and spirochetes, but not found in other bacteria. The 2 a.a. insert in this case is postulated to have been introduced in a common ancestor of the above groups of bacteria as shown by the position of this signature in Fig. 1. The presence of this insert in *T. maritima* constitutes an exception and it may have resulted from lateral gene transfer. The dashes in the alignment show identity with the amino acid in the top line. The asterisks (*) denote unpublished sequences retrived from microbial genome database (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/genom_table_cgi).

exception to the predicted pattern was observed. The observed results provide strong evidence that the signature approach provides a reliable and internally highly consistent means for determining the relative branching orders of bacterial groups.

An important point to be noted from the results shown in Fig. 1 is that the phylogenetic placements of various bacterial species into different groups based on signature sequences show an excellent correlation to that based on the 16S rRNA. Of the 60 bacterial species whose genomes have been sequenced, nearly all of them were assigned to the same groups by both methods. The only possible exception consisted of *Clostridium* species (*C. acetobutylicum* and *C. perfrigens*), which are assigned to the low G+C Gram-positive group based on the 16S rRNA (Olsen *et al.*, 1994). The indel data

TABLE I

Distribution of Identified Indels in Proteins from Completed Bacterial Genomes

Protein	Signature description	No.of genomes with protein	Genomes lacking the protein	No. of genomes with indels Expected/found	No. of genomes lacking the indel Expected/found	Exceptions observed
Ribosomal S12 protein	13 a.a. Low G+C signature	60	0	14/14	46/46	0
Hsp70/DnaK	21–23 a.a. G+/G-insert	60	0	38/38	22/22	0
Hsp90	5 a.a. G+/G-insert	32	nm, cc, at, cm, cp, ct dr, tm, mp, mn, mg, uu, ll, sa, sp, sn, ca, li, lm, at, bm, clp	8/8	24/24	0
Glutamate-semialdehyde aminomutase	1 a.a. insert after *Actinobacteria*	38	bu, hi, rp, rc, ml, at, sl, bm, bb, tp, sp, sn, mg, mn, mp, uu, ll, clp, cg	28/28	10/10	
SecF protein	3-4 a.a. deletion after *Actinobacteria*	52	sp, sn, mg, mn, uu, ll, bu	11/11	41/41	0
Hsp60/GroEL	1 a.a. insert after *Deinococcus*	58	mp, uu	37/37	21/21	0
FtsZ protein	1 a.a. insert after cyanobacteria	52	ct, cp, cm, mn, mg, uu	30/30	22/22	0
Rho protein	2 a.a. insert before spirochetes	50	ll, sp, sn, mg, mp, mn, uu, sy, ns	35/36	15/14	1 (tm)
Ala-tRNA synthetase	4 a.a. after spirochetes	60	0	33/33	27/27	0
Inorganic pyrophosphatase	2 a.a. insert common to *Aquifex* and proteo	43	sa, sp, sn, bs, ll, bp tp, dr, tm, ca, clp, fn, li,lm	28/28	15/15	0
Hsp70/DnaK	2 a.a. proteo insert	60	0	27/27	33/33	0
CTP synthetase	10 a.a. insert before proteobacteria	53	mp, mg, uu, mn, li, lm, fn	27/27	26/26	0
Lon protease	1 a.a. deletion in $\alpha\beta\gamma$-proteobacteria	46	ll, mt, ml, sa, sp, sn, sy, ns, cg	22/22	24/24	0
SecA protein	7 a.a. insert in $\alpha\beta\gamma$-proteobact.	60	0	25/25	35/35	0
HSP70/DnaK	4 a.a. $\beta\gamma$-insert	60	0	16/16	44/44	0
PRPP synthetase	1 a.a. $\beta\gamma$-insert	54	cp, ct, cm, rp	16/16	38/38	0
Biotin carboxylase	1 a.a. $\beta\gamma$-insert	53	mg, mn, mp, uu, bb, tp, tm, bu, fn	15/15	38/38	0
PAC-transformylase	2 a.a. γ-proteo deletion	45	bb, cp, cm, ct, hp, rc, mg, tp, uu, rp	32/32	13/13	0

Note: The abbreviations used in species name are: at, *Agrobacterium tumefaciens*; bb, *Borrelia burgdorferi*; bm, *Brucella melitensis*; bu, *Buchnera sp*., cc, *Caulobacter crescentus*, ca, *Clostridium acetobutylicum*; clp, *Clostridium perfringens*; cm, *Chlamydia muridarum*; cp, *Chlamydophila pneumoniae*; ct, *Chlamydia trachomatis*; dr, *Deinococcus radiodurans*; fn, *Fusobacterium nucleatum*; hp, *Helicobacter pylori*; li, *Listeria innocua*; ll, *Lactococcus lactis*; lm, *Listeria monocytogenes*; mg, *Mycoplasma genitalium*; mn, *M. pneumoniae*; mp, *M. pulmonis*; ns, *Nostoc sp*.; rc, *Rickettsia conorii*; rp, *R. prowazekii*; sa, *Staphylococcus aureus*; sn, *Streptococcus pneumoniae*; sp, *S. pyogenes*; Sy, *Synechococcus* sp. PCC6803; tp, *Treponema pallidum*; uu, *Ureaplasma urealyticum*.

also place these species within the Gram-positive bacteria, but suggest a distinct branching of this group of species from other low G+C Gram-positive bacteria (i.e., *Firmicutes*). Based on the presently available signatures, the *Clostridium* species branch in a similar position as *T. maritima*, however, this does not necessarily mean a specific relationship between these groups. Numerous other species for whom partial genome sequence information is available are also assigned to the same groups using the signature approach and the 16S rRNA trees (unpublished data).

The usefulness of the signature sequence approach for determining the phylogenetic placement and branching order of a given group of species is illustrated by the example of *Fusobacterium nucleatum*, whose genome was recently sequenced (Kapatral *et al*., 2002). Based on its Gram-staining characteristic, *F. nucleatum* is considered to be a Gram-negative bacteria (see Kapatral *et al*., 2002). However, phylogenetic studies based on 16S rRNA as well as several features of its core metabolism, suggest a close relationship of *F. nucleatum* to the *Clostridium* group of species (Jalava and Eerola, 1999; Kapatral *et al*., 2002). In the recent Bergey's manual, this group of bacteria are placed in a separate phylum or division of their own (*Fusobacteria*), unrelated to any of the another groups (Ludwig and Klenk, 2001). However, based on signature sequences found in various proteins, *F. nucleatum* can now be confidently and reliably placed in the same position as *Clostridium* species, which is consistent with its core metabolism as well as branching in the rRNA trees. Thus, it makes no sense at present to assign this group of species a separate phylum or division status. In another instance, the signature sequence approach was used to place *Fibrobacter,* a group which has again been assigned a separate phylum status (Ludwig and Klenk, 2001), to a similar position as the chlamydiae and the *Cytophaga–Flavobacteria–Bacteroides* groups (Griffiths and Gupta, 2001).

The evolutionary scheme based on signature sequences, in addition to its high degree of internal consistency, is also consistent with the major morphological distinctions seen within *Bacteria*. The lack of concordance between phylogenetic inference and the cell ultrastructural characteristics of bacteria was a problem in the past (Stanier *et al*., 1976; Murray, 1986a; Woese, 1992). The model shows that the bacterial groups surrounded by a single membrane (i.e., Gram-positive or monoderm bacteria) are phylogenetically distinct from those surrounded by both an inner and outer cell membrane and separated by a periplasmic compartment (i.e., all true Gram-negative bacteria or diderm bacteria). Of these two groups, the monoderm bacteria are indicated to be ancestral. The deduced scheme places *Deinococcus–Thermus* group of species in an intermediate position between these two groups (Fig. 1). This placement is consistent with the unique characteristics of *Deinococcus*, which contains a thick peptidoglycan layer and shows positive Gram-staining, but it is surrounded by both inner and outer cell membranes, similar to various true Gram-negative bacteria (Murray, 1986b).

ARE THE INFERENCES DERIVED FROM THE INDEL MODEL AFFECTED BY LATERAL GENE TRANSFERS?

The results presented in Table I not only give confidence in the reliability of signatures as a powerful tool for deducing bacterial phylogeny, but they also shed light on another important issue, i.e., of lateral gene transfer (LGT) among bacterial species. The analyses of genomic sequences by means of BLAST searches and other methods have recently led to the proposals that lateral gene transfers are rampant among different bacterial groups, and that such events were particularly pervasive at the earlier stages in bacterial evolution (Woese, 1998; Jain *et al*., 1999; Doolittle, 1999; Campbell, 2000; Koonin *et al*., 2001). This assumption forms the basis of the currently widespread notion that most of the main groups within *Bacteria* are melting pots of genes from different sources, making it very difficult to draw any reliable phylogenetic inferences concerning their branching orders or evolution from a common ancestor (Woese, 1998; Doolittle, 1999; Ludwig and Klenk, 2001). However, this notion of widespread and indiscriminate LGTs among bacterial species is strongly challenged and negated by the results presented here. As seen, different indels in a large number of highly conserved proteins, which have been introduced at different divergence points along the main line of descent, give a highly consistent and reliable picture concerning the phylogenetic placement of different species and their relative branching orders. Of the 936 observations measuring the presence or absence of these indels in various species, only one exception to the predicted pattern was observed. These results strongly argue that the genes containing these indels, most of which are for highly conserved housekeeping functions, have not been affected or corrupted by factors such as LGTs. To account for these results by LGTs, one would have to postulate highly specific LGT events, differing for each indel and involving all species from a number of different bacterial groups. Such a scenario is highly improbable. Likewise, the probability that the observed indels in various proteins have been introduced independently in different species is also astronomically low (Gupta, 2000b).

Lateral gene transfer, however, does occur in bacteria and many such events are readily identified by means of signature sequences. For example, *B. burgdorferi* contains two different Hsp70 homologs, one of which contains the large 21–23 a.a. insert which is distinctive of Gram-negative bacteria, whereas the other does not

contain this insert and thus is similar to those found in Gram-positive bacteria (Gupta, 2001). The BLAST searches and phylogenetic analysis show that this latter homolog is indeed closely related to Gram-positive bacteria (unpublished results) and it is very likely acquired from this group by means of LGT. The presence of this latter homolog in *B. burgdorferi*, which is inconsistent with all of the other signatures concerning the placement of this species, and which is readily identified as derived by means of LGT, does not in any way confuse or affect the phylogenetic placement of this species. Another example of a LGT event between different groups of species is provided by a signature that is present in the protein UDP-*N*-acetylglucosamine pyrophosphorylase (glmU) (Fig. 3). This protein contains a 17 a.a. indel in a conserved region that is commonly shared by various archaebacteria and chlamydiae species. Although this protein is found in most other bacterial groups, they do not contain this insert. The presence of this uniquely shared indel between chlamydiae and archaebacteria suggests a specific relationship between these groups, exclusive of all other bacterial phyla. However, such a relationship is inconsistent with all of the other signatures in different proteins (Fig. 1) as well as various other characteristics distinguishing *Archaea* and *Bacteria*. To account for the presence of this shared signature, the most likely explanation is that this gene was laterally transferred

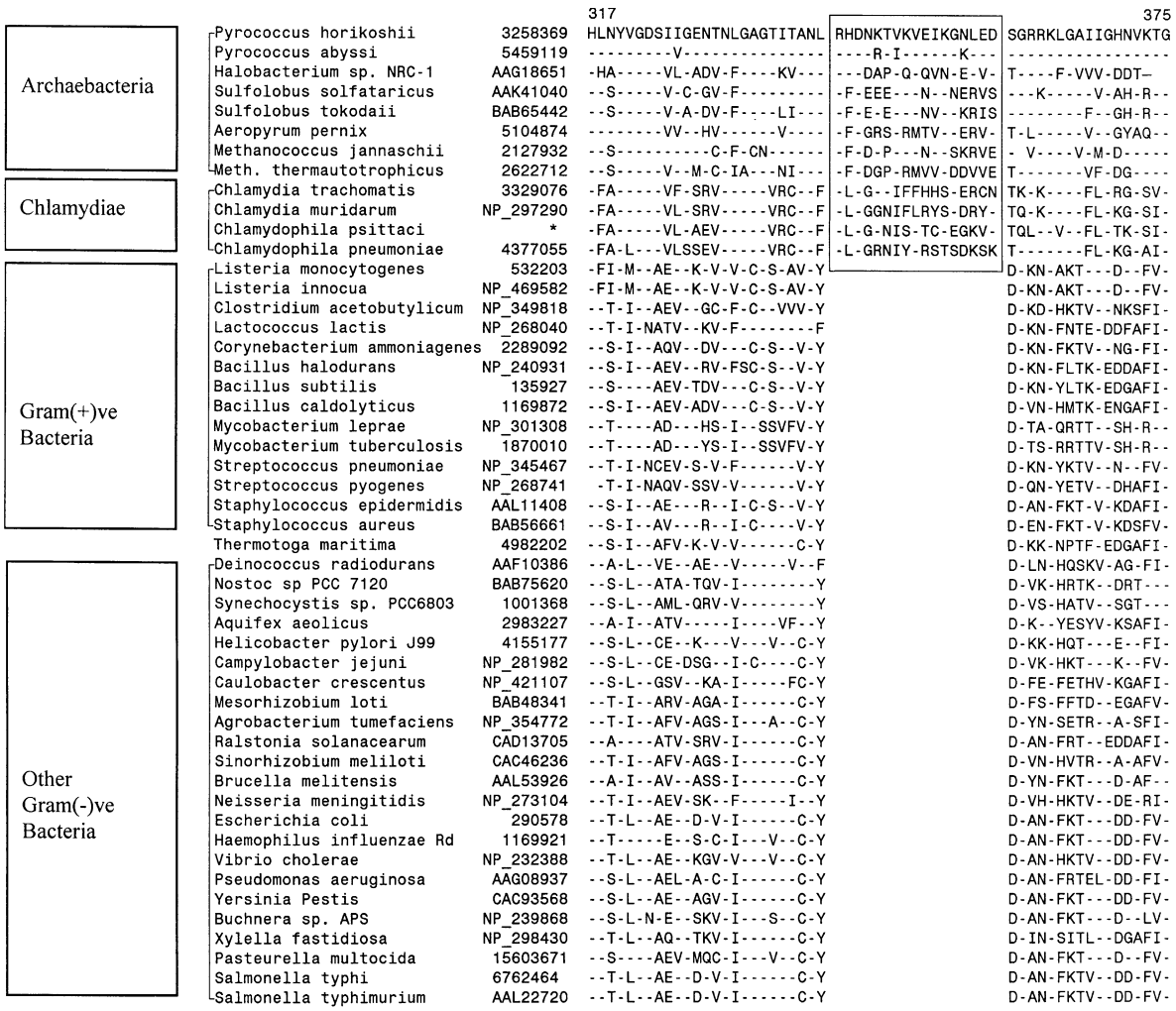

FIG. 3. Partial sequence alignment of glmU protein (UDP-*N*-acetylglucosamine pyrophosphorylase) showing the presence of a 17 a.a. indel that is uniquely shared by various homologs from archaebacteria and the chlamydiae groups of species. The gene for this protein is postulated to have been laterally transferred from archaebacteria to a common ancestor of the chlamydiae species. The dashes in the alignment show identity with the amino acid in the top line.

from an archaebacteria to a common ancestor of the chlamydiae group of species. This inference is supported by the fact that in phylogenetic trees based on this protein, chlamydiae species are indicated to be the closest relatives of archaebacteria (results not shown). Another example of a shared signature that likely results from LGT is found in the protein UDP-*N*-acetylglucosamine 1 -carboxyvinyltransferase (MurA), where a 16 a.a. insert is uniquely present in the chlamydiae homologs and those from the *Streptomyces* species. In phylogenetic trees based on this protein, *Streptomyces* branch with high affinity with the chlamydiae group rather than with other Gram-positive bacteria. These results strongly indicate that the gene for this protein has been laterally transferred between these two groups (Griffiths and Gupta, 2002).

COMING TO GRIP WITH THE CRITICAL ISSUES IN BACTERIAL PHYLOGENY

The main significance of the indel model lies in its ability to provide simple and reliable means to resolve the various important issues in bacterial phylogeny. As shown in Fig. 1, based simply on the presence or absence of various identified signatures, a large number of groups within *Bacteria* can now be defined and distinguished from each other in clear molecular terms. The identified groups correspond to most of the major groups which are presently recognized on the basis of their branching in the 16S rRNA trees (Olsen *et al*., 1994; Ludwig and Klenk, 2001). By simply examining the presence or absence of these signatures, any given or unknown bacterial species could be assigned to any of these groups with a high degree of confidence (Griffiths and Gupta, 2001). A flow chart based on indels which can be used for such purposes has been described in earlier work (Gupta, 2000b). In addition to the signatures shown in Fig. 1, we have also identified a large number of signatures that are specific for particular groups such as cyanobacteria, chlamydiae, *Deinococcus–Thermus*, spirochetes, or α-proteobacteria, which also serve to identify and define these groups (Griffiths and Gupta, 2002 and unpublished results). The defining of these groups can be done with greater degree of confidence using the signature approach than with the traditional tree building method, due to the lack of subjectivity involved in the interpretation of the results. The assignment of species to these groups by this method is based simply on the presence or absence of different indels and it does not require any additional assumption or estimation such as establishing a degree of confidence at each node of a phylogenetic tree.

In our work, we have adopted a definition of the main divisions as those groups of species whose branching order from the main line of descent can be clearly established, and which is distinct from all other identified main groups (Gupta, 1998a,2000b,2001). Thus, in our scheme the proteobacteria have been divided into a number of main groups, each of which can be clearly distinguished from the other and their relative branching order can be clearly deduced. It has been previously suggested that each of these four proteobacterial groups should be recognized as a division or phylum, similar to the other main groups or divisions among *Bacteria* (Gupta, 2000b). In contrast, a number of other groups such as chlamydiae, Cytophaga–flavobacteria–bacteroides (CFB group) and green sulfur bacteria are not assigned separate main group statuses as they presently branch in the same position based on the available signature sequences. Similarly, the $\delta-$ and ε-proteobacteria also branch in the same position. The number of main groups that we have thus far identified using the signature sequence approach represents a minimal number. As additional signature sequences or other information that clarifies their branching order become available, some of these groups will likely be elevated to a separate main group or phylum status. In the mean time, the distinct group of species branching in the same position are considered in the present scheme as subdivisions of the main group (Gupta, 2000b). The division of *Bacteria* into various main groups or subdivisions in the present scheme is based strictly on a genealogical basis, which is the most logical way of understanding phylogeny. Thus, the signature sequence approach has brought us much closer to understanding the evolutionary relationships among bacteria which has been a long sought after goal of microbiology. As additional genome sequences from divergent bacterial species become available, they will provide further means to test and vindicate the proposed model.

ACKNOWLEDGMENTS

This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada and Canadian Institute of Health Research.

REFERENCES

Balows, A., Trüper, H. G., Dworkin, M., Harder, W., and Schleifer, K. H. 1992. "The Prokaryotes," 2nd ed., Springer-Verlag, New York.

Brendel, V., Brocchieri, L., Sandler, S. J., Clark, A. J., and Karlin, S. 1997. Evolutionary comparisons of RecA-like proteins across all major kingdoms of living organisms, *J. Mol. Evol.* **44,** 528–541.

Buchanan, R. E. 1925. "General Systematic Bacteriology," Williams & Wilkins, Baltimore.

Buchanan, R. E., and Gibbons, N. E. 1974. "Bergey's Manual of Determinative Bacteriology," 8th ed., Williams & Wilkins, Baltimore.

Campbell, A. M. 2000. Lateral gene transfer in prokaryotes, *Theor. Popul. Biol.* **57,** 71–77.

Doolittle, W. F. 1999. Phylogenetic classification and the universal tree, *Science* **284,** 2124–2128.

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.

Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N., and Woese, C. R. 1980. The phylogeny of prokaryotes, *Science* **209,** 457–463.

Griffiths, E., and Gupta, R. S. 2001. The use of signature sequences in different proteins to determine the relative branching order of bacterial divisions: Evidence that *Fibrobacter* diverged at a similar time to *Chlamydia* and the *Cytophaga–Flavobacterium–Bacteroides* division, *Microbiology* **147,** 2611–2622.

Griffiths, E., and Gupta, R. S. 2002. Protein signatures distinctive of chlamydial species: Horizontal transfer of cell wall biosynthesis genes *glmU* from Archaebacteria to Chlamydiae, and *murA* between Chlamydiae and *Streptomyces*, *Microbiology* (in press).

Gupta, R. S. 1998a. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes, *Microbiol. Mol. Biol. Rev.* **62,** 1435–1491.

Gupta, R. S. 1998b. 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.

Gupta, R. S. 1998c. 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.

Gupta, R. S. 2000a. The *natural* evolutionary relationships among prokaryotes, *Crit. Rev. Microbiol.* **26,** 111–131.

Gupta, R. S. 2000b. The phylogeny of Proteobacteria: relationships to other eubacterial phyla and eukaryotes, *FEMS Microbiol. Rev.* **24,** 367–402.

Gupta, R. S. 2001. The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins, *Int. Microbiol.*, in press.

Gupta, R. S., and Singh, B. 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.

Holt, J. G. E. 1984. "Bergey's Manual of Systematic Bacteriology," Williams & Wilkins, Baltimore.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes, *Proc. Natl. Acad. Sci. USA* **86,** 9355–9359.

Jain, R., Rivera, M., and Lake, J. A. 1999. Horizontal gene transfer among genomes: The complexity hypothesis, *Proc. Natl. Acad. Sci. USA* **96,** 3801–3806.

Jalava, J., and Eerola, E. 1999. Phylogenetic analysis of *Fusobacterium alocis* and *Fusobacterium sulci* based on 16S rRNA gene sequences: Proposal of *Filifactor alocis* (Cato, Moore and Moore) comb. nov. and *Eubacterium sulci* (Cato, Moore and Moore) comb. nov., *Int. J. Syst. Bacteriol.* **49,** 1375–1379.

Kapatral, V., Anderson, I., Ivanova, N., Reznik, G., Los, T., Lykidis, A., Bhattacharyya, A., Bartman, A., Gardner, W., Grechkin, G., Zhu, L., Vasieva, O., Chu, L., Kogan, Y., Chaga, O., Goltsman, E., Bernal, A., Larsen, N., D'Souza, M., Walunas, T., Pusch, G., Haselkorn, R., Fonstein, M., Kyrpides, N., and Overbeek, R. 2002. Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586, *J. Bacteriol.* **184,** 2005–2018.

Karlin, S., Weinstock, G. M., and Brendel, V. 1995. Bacterial classifications derived from recA protein sequence comparisons, *J. Bacteriol.* **177**, 6881–6893.

Kluyver, A. J., and van Niel, C. B. 1936. Prospects for a natural system of classification of bacteria, *Zentralbl. Bakteriol. Parasitenk. Infektionskr. II* **94,** 369–403.

Koonin, E. V., Makarova, K. S., and Aravind, L. 2001. Horizontal gene transfer in prokaryotes: Quantification and classification, *Annu. Rev. Microbiol.* **55,** 709–742.

Lake, J. A. 1991 . The order of sequence alignment can bias the selection of tree topology, *Mol. Biol. Evol.* **8,** 378–385.

Ludwig, W., and Klenk, H.-P. 2001. Overview: A phylogenetic backbone and taxonomic framework for prokaryotic systematics. *in* "Bergey's Manual of Systematic Bacteriology" (D. R. Boone, and R. W. Castenholz, Eds.), pp. 49–65, Springer-Verlag, Berlin.

Ludwig, W., and Schleifer, K. H. 1999. Phylogeny of *Bacteria* beyond the 16S rRNA standard, *ASM News* **65**, 752–757.

Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T. Jr., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M., and Tiedje, J. M. 2001. The RDP-II (ribosomal database project), *Nucleic Acids Res.* **29**, 173–174.

Morowitz, H. J. 1992. "Beginnings of Cellular Life: Metabolism Recapitulates Biogenesis," pp. 1–195, Yale Univ. Press, New Haven, CT.

Murray, R. G. E. 1986a. The higher taxa, or, a place for everything...? *in* "Bergey's Manual of Systematic Bacteriology?" (P. H. A. Sneath, N. S. Mair, M. E. Sharpe, and J. G. Holt), pp. 31–34, Williams & Wilkins, Baltimore.

Murray, R. G. E. 1986b. Family II. *Deinococcaceae* Brooks and Murray 1981, 356^{VP} *in* "Bergey's Manual of Systematic Bacteriology" (P. H. A. Sneath, N. S. Mair, M. E. Sharpe, and J. G. Holt, Eds.), pp. 1035–1043, Williams & Wilkins, Baltimore.

Olsen, G. J., and Woese, C. R. 1993. Ribosomal RNA: A key to phylogeny, *FASEB J.* **7,** 113–123.

Olsen, G. J., Woese, C. R., and Overbeek, R. 1994. The winds of (evolutionary) change: Breathing new life into microbiology, *J. Bacteriol.* **176,** 1–6.

Orla-Jensen, S. 1909. Die Hauptlinien de natürlichen Baketerien-systems nebst einer Uebersicht der Gärungsphenomene, *Zentralbl. Bakteriol. Parasitenk. Infektionskr. II* **22,** 305–346.

Stackebrandt, E., Murray, R. G. E., and Trüper, H. G. 1988. *Proteobacteria classis nov*., a name for the phylogenetic taxon that includes the "Purple bacteria and their Relatives," *Int. J. Syst. Bacteriol.* **38**, 321–325.

Stanier, R. Y. 1941. The main outlines of bacterial classification, *J. Bacteriol.* **42,** 437–466.

Stanier, R. Y., Adelberg, E. A., and Ingraham, J. L. 1976. "The Microbial World," 4th ed., pp. 1–871, Prentice-Hall Inc., Engelwood Cliffs, NJ.

Stanier, R. Y., and van Niel, C. B. 1962. The concept of a bacterium, *Arch. Mikrobiol.* **42,** 17–35.

Viale, A. M., Arakaki, A. K., Soncini, F. C., and Ferreyra, R. G. 1994. Evolutionary relationships among eubacterial groups as inferred from GroEL (chaperonin) sequence comparisons, *Int. J. Syst. Bacteriol.* **44,** 527–533.

Woese, C. R. 1987. Bacterial evolution, *Microbiol. Rev.* **51,** 221–271.

Woese, C. R., 1991. The use of ribosomal RNA in reconstructing evolutionary relationships among bacteria. *in* "Evolution at Molecular Level" (R. K. Selander, A. G. Clark, and T. S. Whittmay, Eds.), pp. 1–24, Sinauer Associates Inc., Sunderland, MA.

Woese, C. R., 1992. Prokaryote systematics: The evolution of a science. *in* "The Prokaryotes" (A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K. H. Schleifer, Eds.), pp. 3–18, Springer-Verlag, New York.

Woese, C. R. 1998. The universal ancestor, *Proc. Natl. Acad. Sci. USA* **95,** 6854–6859.

Woese, C. R., Stackebrandt, E., Macke. R. J., and Fox, G. E. 1985. A phylogenetic definition of the major eubacterial taxa, *Syst. Appl. Microbiol.* **6,** 143–151.

Zuckerkandl, E., and Pauling, L. 1965. Molecules as documents of evolutionary history, *J. Theor. Biol.* **8,** 357–366.