Int Microbiol (2001) 4: 187–202
DOI 10.1007/s10123-001-0037-9

REVIEW ARTICLE

Radhey S. Gupta

The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins

Received: 3 August 2001 / Accepted: 5 September 2001 / Published online: 20 December 2001

Abstract The presence of shared conserved inserts and deletions (indels or signature sequences) in proteins provides a powerful means for understanding the evolutionary relationships among the Bacteria. Using such indels, all of the main groups within the Bacteria can be defined in clear molecular terms and it has become possible to deduce that they branched from a common ancestor in the following order: Low G+C Gram-positive → High G+C Gram-positive → *Deinococcus–Thermus* → Cyanobacteria → Spirochetes → *Aquifex–Chlamydia–Cytophaga* → Proteobacteria-1 (ϵ, δ) → Proteobacteria-2 (α)→ Proteobacteria-3 (β) → Proteobacteria -4 (γ). The usefulness of this approach for understanding bacterial phylogeny was examined here using sequence data from various completed bacterial genomes. By using 12 indels in highly conserved and widely represented proteins, the species from all 41 completed bacterial genomes were assigned to different groups; and the observed distribution of these indels in different species was then compared with that predicted by the signature sequence model. The presence or absence of these indels in various proteins in different bacteria followed the pattern exactly as predicted; and, in more than 450 observations, no exceptions or contradictions in the placement of indels were observed. These results provide strong evidence that lateral gene transfer events have not affected the genes containing these indels to any significant extent. The phylogenetic placement of bacteria into different groups based on signature sequences also showed an excellent correlation with the 16 S rRNA with 39 of the 41 species assigned to the same group by both methods. These results strongly vindicate the usefulness of the signature sequence approach to understanding phylogeny within the Bacteria and show that it provides a reliable and internally consistent means for the placement of bacterial species into different groups and for determining the relative branching order of the groups.

Keywords Indels · Signature sequences · Bacterial genomes · Lateral gene transfer · Phylogeny

R.S. Gupta
Department of Biochemistry,
McMaster University,
Hamilton L8N 3Z5, Ontario, Canada
E-mail: gupta@mcmaster.ca
Tel.: +1-905-5259140
Fax: +1-905-5229033

Introduction

Our current understanding of evolutionary relationships among the Bacteria, which comprise the vast majority of the known prokaryotes, is almost entirely based on the 16 S rRNA sequences [4, 40, 51]. Based on oligonucleotide signatures and the branching pattern of bacteria in the 16 S rRNA trees, 11 main groups (or divisions) among the Bacteria were originally proposed [69, 70, 72]. These included: *Thermotogales*, green nonsulfur bacteria, *Deinococci* and relatives, *Spirochetes*, green sulfur bacteria, Cyanobacteria, Gram-positive bacteria, purple bacteria and relatives (*Proteobacteria*), *Bacteriodes–Flavobacteria–Cytophaga* and relatives, *Planctomyces* and relatives, and chlamydiae. At the time when these divisions were proposed, the rRNA sequence database was quite limited and clear distinctions between these groups was possible on the basis of oligonucleotide signatures or long "naked" branches that separated these groups in the trees [69, 72]. However, in the past 15–20 years, as the sequence database for rRNA has rapidly expanded [42], distinguishing between these divisions on the basis of either of these criteria has become increasingly difficult and imprecise [40, 41]. In recent years, in addition to the above groups, many additional groups or divisions within the Bacteria have been suggested (i.e., *Aquificales*, *Desulfurobacterium*, *Dictyoglomus*, *Fibrobacter*, *Flexistipes*, *Fusobacteria*, *Holophaga*, *Nitrospira*, *Verrucomicrobium*) [40, 41]. In the absence of well defined criteria for the major divisions, it is unclear how many of these newly described groups actually comprise new divisions within the Bacteria. To place the bacterial phylogeny on a firmer base, it is essential to develop clear molecular criteria by which the different major groups (phyla or divisions) within the Bacteria can be defined and distinguished from each other. Another issue central to bacterial phylogeny is to determine how the different main groups or divisions within the Bacteria are related to each other and how they branched from a common ancestor [21]. Such relationships are not resolved in phylogenetic trees based on rRNA or various proteins [6, 11, 40, 51, 69]. This has led to a growing acceptance of the notion that such relationships are unresolvable and that all the main groups within the Bacteria probably branched off simultaneously from the common ancestor [11, 40, 41, 71].

We recently described a new approach that makes use of conserved inserts and deletions (referred to as indels or signature sequences) found in various proteins, which provides valuable information regarding the issues that are not resolved in the rRNA trees [19, 23]. Based simply on the presence or absence of specific signature sequences, all of the major groups within the Bacteria can be clearly defined and distinguished from each other. Further, this approach also permits a logical deduction of the relative branch order of different main groups from a common ancestor [19, 23, 26], which has been a major impediment in understanding bacterial phylogeny. In the past few years, the entire genomes of many bacterial species have been sequenced, representing all major groups within the Bacteria (http://www. ncbi.nlm.nih.gov:80/PMGifs/Genomes/micr.html). This provides us with a valuable means to test in an objective manner the usefulness and validity of the signature sequence approach for determining the phylogenetic placement and branching order of the bacterial species. Results of these studies presented here strongly evidence that this approach provides a reliable and internally consistent means for the phylogenetic placement of species into different groups and for determining their relative branching order. Importantly, the assignment of bacterial species into different groups using this new approach shows a very high degree of correlation to that based on the 16 S rRNA trees. Therefore, this new approach is not contradictory to the 16 S rRNA analyses but complements the latter studies in important respects, by providing information regarding issues that are not resolved in such phylogenies.

Results and Discussion

Bacterial genomes and signature sequence

The information for various bacterial species whose complete genomes have been sequenced to date is given in Table 1. The sequence information is presently available for 41 bacterial genomes, representing all of the main groups within the Bacteria including: α-, β-, γ-, and ϵ-Proteobacteria, *Aquificales*, *Chlamydia*, Cyanobacteria, *Deinococcus–Thermus* group, Spirochetes, *Thermotoga,* several members of the low G + C Gram-positive bacteria including the mycoplasmas, and high G + C Gram-positive species (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html). In our earlier work, a large number of sequence signatures in different proteins were identified [19, 23]. Some of these signatures are specific for the particular groups and they provide no information regarding relationships to other groups [19]. However, of the identified signatures, a group of 12 signatures has proven most useful for distinguishing between the major groups within the Bacteria and for determining their branch order (Fig. 1). The sequence information for these signatures for various bacterial species whose genomes have been sequenced was obtained by basic local alignment search tool (BLAST) searches (http://www.ncbi.nlm.nih.gov) on either the non-redundant database or on individual genome sequences.

The rationale for using conserved indels for phylogenetic studies

The rationale for using conserved indels for evolutionary studies has been discussed in detail in earlier work [19, 20]. When a conserved indel of defined length and sequence (referred to as a signature sequence) that is flanked by conserved regions to ensure its reliability is found at the same position within a given protein (or gene) from different species, then 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 [19, 56]. Thus, based on the presence or absence of a signature, the species containing or lacking the signature can be divided into two unambiguous groups. The well defined indels in different genes/proteins also provide useful milestones for evolutionary events, since all species emerging from the ancestral cell in which a given indel was first introduced are expected to contain the indel, whereas all species that existed prior to this event or which did not evolve from this ancestor will not contain the indel [19, 20]. Thus, by using well defined indels in proteins that were introduced at various stages in evolutionary history, it should be possible to deduce the branching order of different groups of species from a common ancestor.

Testing the signature sequence model using completed bacterial genomes

Figure 1 shows the signature sequences in proteins that have proven most useful for distinguishing the major groups within the Bacteria and to determine their relative

Table 1 Details of bacterial species whose genomes have been sequenced

Bacterial species	Accession number	Bacterial group/division	Reference
Aquifex aeolicus	NC000918	Aquificales	[10]
Bacillus halodurans C-125	NC002570	Low G + C Gram-positive	[65]
B. subtilis	NC000964	Low G + C Gram-positive	[36]
Borrelia burgdorferi	NC001318	Spirochaetales	[14]
Buchnera sp. APS	NC002528	γ-Proteobacteria	[59]
Campylobacter jejuni	NC002163	ϵ-Proteobacteria	[53]
Caulobacter crescentus	NC002696	α-Proteobacteria	[50]
Chlamydia muridarum	NC002182	Chlamydiales	[55]
C. trachomatis	NC000117	Chlamydiales	[62]
Chlamydophila pneumoniae CWL029	NC000922	Chlamydiales	[31]
C. pneumoniae AR39	NC002179	Chlamydiales	[55]
C. pneumoniae J138	NC002491	Chlamydiales	[60]
Deinococcus radiodurans	NC001263	*Deinococcus/Thermus*	[68]
Escherichia coli K12	NC000913	γ-Proteobacteria	[5]
E. coli OI57:H7 EDL933	NC002655	γ-Proteobacteria	[54]
E. coli OI57:H7	NC002695	γ-Proteobacteria	[43]
Haemophilus influenzae	NC000907	γ-Proteobacteria	[13]
Helicobacter pylori 26695	NC00915	ϵ-Proteobacteria	[67]
H. pylori J99	NC000921	ϵ-Proteobacteria	[1]
Lactococcus lactis	NC002662	Low G + C Gram-positive	(unpublished)
Mesorhizobium loti	NC002678	α-Proteobacteria	[32]
Mycoplasma genitalium	NC000908	Low G + C Gram-positive	[15]
Mycobacterium leprae	NC002677	High G + C Gram-positive	[9]
M. tuberculosis H37Rv	N000962	High G + C Gram-positive	[8]
M. tuberculosis CDC1551	NC002755	High G + C Gram-positive	(unpublished)
M. pneumoniae	NC000912	Low G + C Gram-positive	[28]
M. pulmonis	NC002771	Low G + C Gram-positive	[7]
Neisseria meningitidis MC58	NC002183	β-Proteobacteria	[66]
N. meningitidis Z2491	NC002263	β-Proteobacteria	[52]
Pasteurella multocida	NC002663	γ-Proteobacteria	[44]
Pseudomonas aeruginosa	NC002516	γ-Proteobacteria	[64]
Rickettsia prowazekii	NC000963	α-Proteobacteria	[2]
Staphylococcus aureus N315	NC002795	Low G + C Gram-positive	[37]
S. aureus Mu50	NC002758	Low G + C Gram-positive	[37]
Streptococcus pyogenes	NC002737	Low G + C Gram-positive	[12]
Synechocystis sp. PCC6803	NC000911	Cyanobacteria	[33]
Thermotoga maritima	NC000853	Thermotogales	[49]
Treponema pallidum	NC000919	Spirochaetales	[16]
Ureaplasma urealyticum	NC002162	Low G + C Gram-positive	[17]
Vibrio cholerae	NC002506	γ-Proteobacteria	[27]
Xylella fastidiosa	NC002488	γ-Proteobacteria	[61]

branch orders. Based upon our analyses, these signatures have been introduced in these proteins at the indicated stages of the evolution of the bacterial groups. Hence, by using them, it should be possible to assign any given bacterial species into one of these groups and to determine its branching order, relative to the other groups.

To test in an objective manner the validity of the evolutionary model based on these signatures, we have analyzed the sequence data from various completed bacterial genomes using this approach. For these purposes, an alignment of the corresponding proteins from bacterial species whose complete genomes have been sequenced was carried out; and the presence or absence of the indicated signatures was determined. This information was then used for the phylogenetic placement of the species into different groups and to determine whether the distribution of these signatures in different species followed the pattern, as predicted by the model, or whether the results obtained were more readily explained by other mechanisms, such as either independent occurrence of the indels in different species, or lateral gene transfer (LGT) between species.

According to the model, once an indel has been introduced in an ancestral lineage, various groups of species emerging after that point should all contain the indel, whereas all species from different groups that existed prior to the introduction of the indel should lack the indel. However, if such indels have been introduced either independently in various species or if the genes containing these indels have been frequently horizontally transferred from one species to another, then the presence or absence of these indels 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. Thus, by determining how closely the results of the indel data follow the predictions of the model and how many exceptions to this are observed, it should be possible to objectively determine whether the inferences based on these indels are reliable and to what extent they

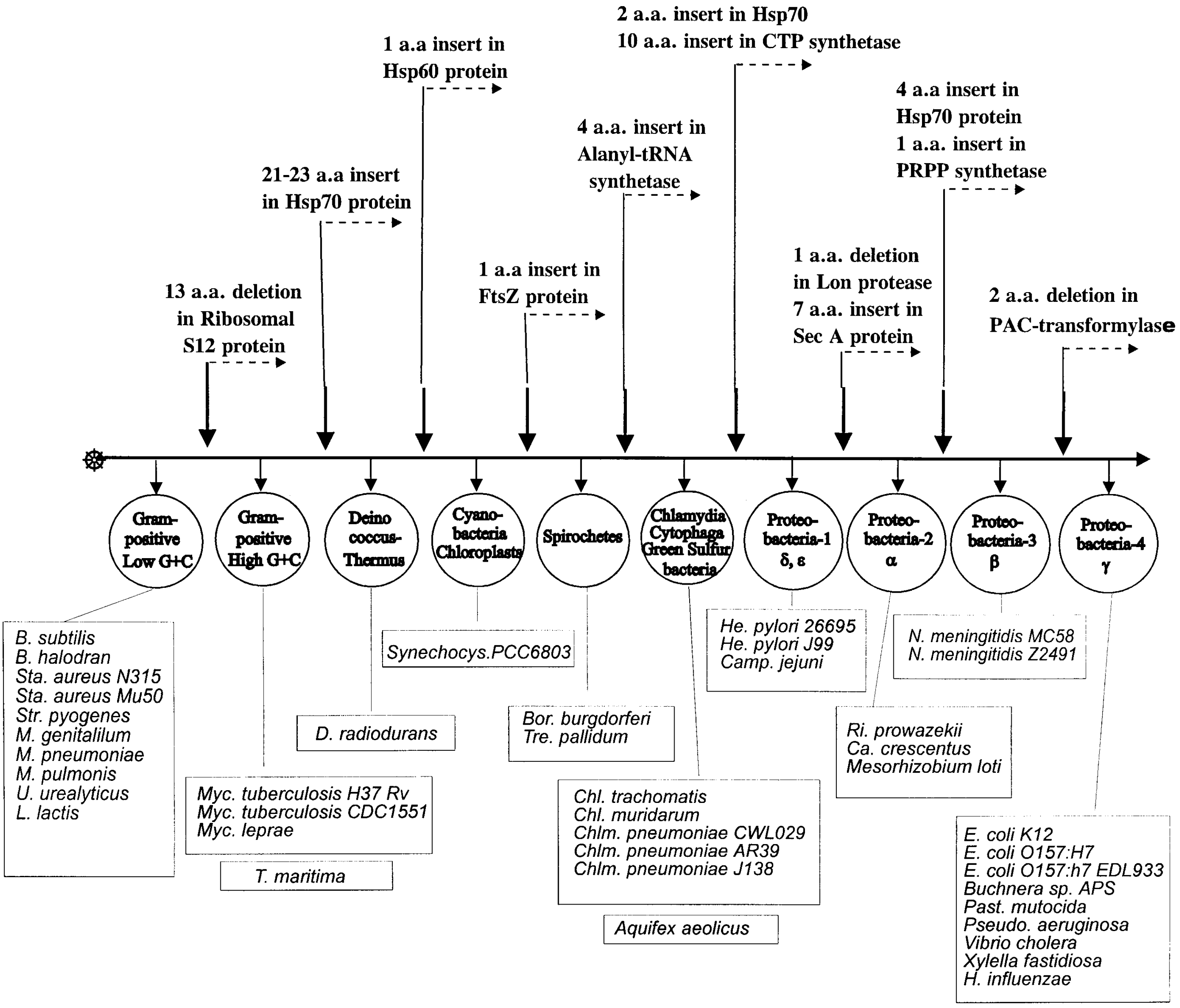


Fig. 1 Phylogenetic placement and relative branching order of bacterial species from completed genomes, based on the indel model developed in earlier work [19, 23]. The *arrows above the line* indicate the specific stages where the indicated signatures in various proteins have been introduced. The model predicts that all bacterial groups *to the right* of these arrows should contain the indicated signatures whereas all groups *to the left* should lack them. The sequences from various bacterial genome conform to the expected patterns, with no exceptions observed. The phylogenetic assignment of bacterial species whose genomes have been sequenced into different groups based on these signatures is *indicated below the line*

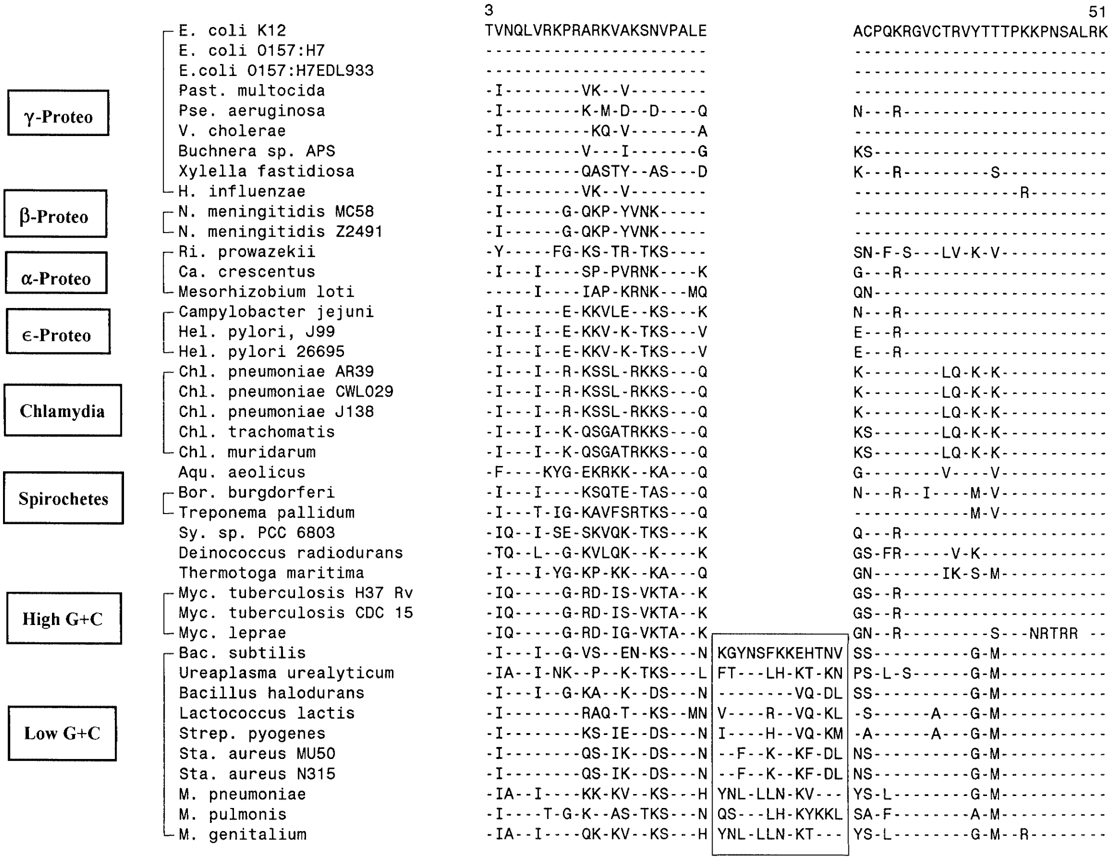
have been corrupted by other factors. The results for these signatures for the bacterial species whose genomes have been sequenced are discussed below.

Ribosomal S12 protein

Ribosomal S12 protein is an essential protein found in all sequenced microbial genomes. A 13-amino-acid indel in a highly conserved region of this protein has been shown to distinguish the low G + C Gram-positive bacteria from all other bacteria [19, 20]. Among the completed microbial genomes, this indel was present in all of the low G + C Gram-positive species, i.e. *Bacillus subtilis*, *B. halodurans*, *Lactococcus lactis*, *Mycoplasma genitalium*, *M. pneumoniae*, *Staphylococcus aureus* (N315, MU50 strains), *Streptococcus pyogenes*, and *Ureaplasma urealyticus*, but not in any other bacteria (Fig. 2, see Appendix). Thus, as indicated in Fig. 1, this signature is a distinctive characteristic of the low G + C Gram-positive group and, based upon it, the species belonging to this group can be clearly distinguished from all other bacteria.

Hsp70 protein

The Hsp70/DnaK family of proteins, which carry out an essential molecular chaperone function in protein-folding

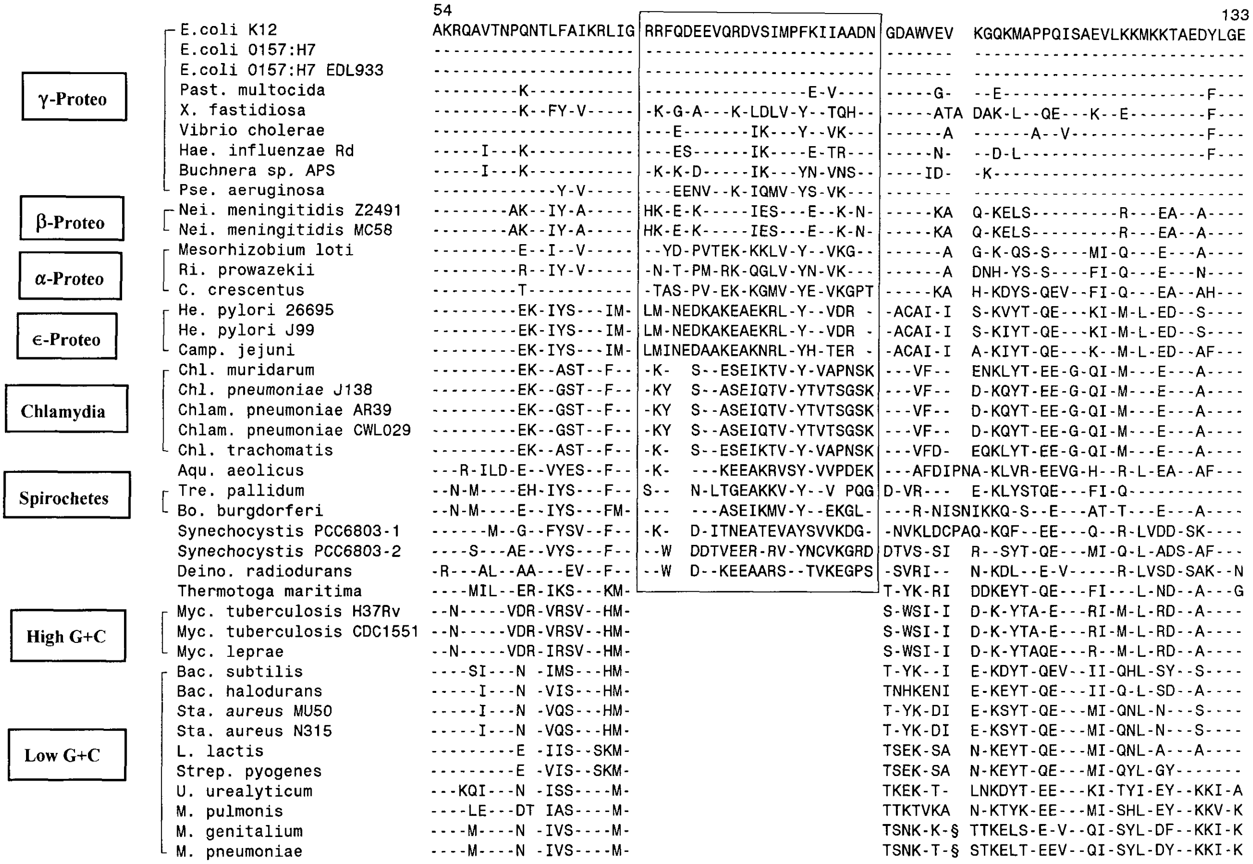


```
                             3                                                   51
E. coli K12                  TVNQLVRKPRARKVAKSNVPALE               ACPQKRGVCTRVYTTTPKKPNSALRK
E. coli O157:H7              -----------------------               --------------------------
E.coli O157:H7EDL933         -----------------------               --------------------------
Past. multocida              -I--------VK--V--------               --------------------------
Pse. aeruginosa              -I--------K-M-D--D----Q               N---R---------------------
V. cholerae                  -I---------KQ-V-------A               --------------------------
Buchnera sp. APS             ----------V---I-------G               KS------------------------
Xylella fastidiosa           -I--------QASTY--AS---D               K---R---------S-----------
H. influenzae                -I--------VK--V--------               -----------------R--------
N. meningitidis MC58         -I------G-QKP-YVNK-----               --------------------------
N. meningitidis Z2491        -I------G-QKP-YVNK-----               --------------------------
Ri. prowazekii               -Y-----FG-KS-TR-TKS----               SN-F-S---LV-K-V-----------
Ca. crescentus               -I---I----SP-PVRNK----K               G---R---------------------
Mesorhizobium loti           -----I----IAP-KRNK---MQ               QN------------------------
Campylobacter jejuni         -I------E-KKVLE--KS---K               N---R---------------------
Hel. pylori, J99             -I---I--E-KKV-K-TKS---V               E---R---------------------
Hel. pylori 26695            -I---I--E-KKV-K-TKS---V               E---R---------------------
Chl. pneumoniae AR39         -I---I--R-KSSL-RKKS---Q               K--------LQ-K-K-----------
Chl. pneumoniae CWL029       -I---I--R-KSSL-RKKS---Q               K--------LQ-K-K-----------
Chl. pneumoniae J138         -I---I--R-KSSL-RKKS---Q               K--------LQ-K-K-----------
Chl. trachomatis             -I---I--K-QSGATRKKS---Q               KS-------LQ-K-K-----------
Chl. muridarum               -I---I--K-QSGATRKKS---Q               KS-------LQ-K-K-----------
Aqu. aeolicus                -F----KYG-EKRKK--KA---Q               G--------V----V-----------
Bor. burgdorferi             -I---I----KSQTE-TAS---Q               N---R--I----M-V-----------
Treponema pallidum           -I---T-IG-KAVFSRTKS---Q               ------------M-V-----------
Sy. sp. PCC 6803             -IQ--I-SE-SKVQK-TKS---K               Q---R---------------------
Deinococcus radiodurans      -TQ--L--G-KVLQK--K----K               GS-FR-----V-K-------------
Thermotoga maritima          -I---I-YG-KP-KK--KA---Q               GN-------IK-S-M-----------
Myc. tuberculosis H37 Rv     -IQ-----G-RD-IS-VKTA--K               GS--R---------------------
Myc. tuberculosis CDC 15     -IQ-----G-RD-IS-VKTA--K               GS--R---------------------
Myc. leprae                  -IQ-----G-RD-IG-VKTA--K               GN--R---------S---NRTRR --
Bac. subtilis                -I---I--G-VS--EN-KS---N KGYNSFKKEHTNV SS----------G-M-----------
Ureaplasma urealyticum       -IA--I-NK--P--K-TKS---L FT---LH-KT-KN PS-L-S------G-M-----------
Bacillus halodurans          -I---I--G-KA--K--DS---N --------VQ-DL SS----------G-M-----------
Lactococcus lactis           -I--------RAQ-T--KS--MN V----R--VQ-KL -S------A---G-M-----------
Strep. pyogenes              -I--------KS-IE--DS---N I----H--VQ-KM -A------A---G-M-----------
Sta. aureus MU50             -I--------QS-IK--DS---N --F--K--KF-DL NS----------G-M-----------
Sta. aureus N315             -I--------QS-IK--DS---N --F--K--KF-DL NS----------G-M-----------
M. pneumoniae                -IA--I----KK-KV--KS---H YNL-LLN-KV--- YS-L--------G-M-----------
M. pulmonis                  -I----T-G-K--AS-TKS---N QS---LH-KYKKL SA-F--------A-M-----------
M. genitalium                -IA--I----QK-KV--KS---H YNL-LLN-KT--- YS-L--------G-M--R--------
```

Fig. 2 Alignment of ribosomal S12 protein sequences from completed bacterial genomes showing a 13-amino-acid insert (*boxed*) that is distinctive of the low G + C Gram-positive bacteria. *Dashes* in all sequence alignments show identity with the amino acid on the top line

and other cellular processes, are found in all completed bacterial genomes. A prominent signature, consisting of an indel of 21–23 amino acids, has been identified in the Hsp70 protein that distinguishes Gram-positive bacteria from Gram-negative bacteria [19, 20, 24]. The large indel in the Hsp70 protein is present in homologues from different Gram-negative bacteria but is absent from those of the Gram-positive bacteria (Fig. 3). The Gram-negative bacteria are defined in our work by the presence of both an inner and outer cell membrane, rather than on the basis of the Gram-staining reaction, which is a variable characteristic [19, 20]. Among the completed genomes, this indel, as expected, was found in all Gram-negative bacteria, but was not present in any of the Gram-positive bacteria, nor was it present in *Thermotoga maritima* or various mycoplasma species, supporting their grouping with the Gram-positive bacteria. In *Synechocystis* sp., multiple homologues for Hsp70 were found [33] and all of these contain the large insert (Fig. 3) [26]. Two different homologues for Hsp70 were also found in the genome of the spirochete species *Borrelia burgdorferi* [14]. One of these homologues, which contained the large insert (GenBank no. 2688438), was closely related to the other spirochete species, *Treponema pallidum*. In contrast, a second Hsp70 homologue in *B. burgdorferi* (GenBank no. 2688201) lacked the large insert. BLAST searches on this homologue indicated that all of the top scores in this case consisted of various Gram-positive bacteria and archaeobacteria. Thus, it is likely that this homologue is derived from Gram-positive bacteria by means of LTG. The Hsp70 sequences are available in the databases for more than 150 bacterial homologues. Of these, this insert is not found in any Gram-positive bacteria and, with the single exception of *B. burgdorferi* noted here, it is a distinctive characteristic of all Gram-negative bacteria [19, 26].

Since the indel in Hsp70 divides the Bacteria into two structurally distinct groups, the question arises whether this indel is an insert in the Gram-negative or a deletion in the Gram-positive. Several lines of evidence support the former of these two possibilities. First, based on the accepted rooting of the prokaryotic tree using duplicated elongation factor EF-1/EF-2 sequences [29], the root of the prokaryotic tree has been shown to lay between archaebacteria and Gram-positive bacteria [19]. The Hsp70 homologues from both these groups of prokaryotes lack this indel, which strongly suggests that this indel is an insert in the Gram-negative bacteria that evolved at a later stage. A second argument supporting this inference is based on the sequence similarity between Hsp70 and another


γ-Proteo
β-Proteo
α-Proteo
ϵ-Proteo
Chlamydia
Spirochetes
High G+C
Low G+C

```
                              54                                                                      133
E.coli K12                    AKRQAVTNPQNTLFAIKRLIG RRFQDEEVQRDVSIMPFKIIAADN GDAWVEV   KGQKMAPPQISAEVLKKMKKTAEDYLGE
E.coli O157:H7                --------------------- ------------------------ -------   ----------------------------
E.coli O157:H7 EDL933         --------------------- ------------------------ -------   ----------------------------
Past. multocida               ---------K----------- -----------------E-V---- -----G-   --E--------------------F---
X. fastidiosa                ---------K--FY-V----- -K-G-A---K-LDLV-Y--TQH-- -----ATA DAK-L--QE---K--E--------F---
Vibrio cholerae               --------------------- ---E-------IK---Y--VK--- ------A   ------A--V-------------F---
Hae. influenzae Rd            -----I---K----------- ---ES------IK----E-TR--- -----N-   --D-L------------------F---
Buchnera sp. APS              -----I---K----------- -K-K-D-----IK---YN-VNS-- ----ID-   -K--------------------------
Pse. aeruginosa               ------------Y-V------ ---EENV--K-IQMV-YS-VK--- -------   ----------------------------
Nei. meningitidis Z2491       --------AK--IY-A----- HK-E-K-----IES---E--K-N- -----KA   Q-KELS---------R---EA--A----
Nei. meningitidis MC58        --------AK--IY-A----- HK-E-K-----IES---E--K-N- -----KA   Q-KELS---------R---EA--A----
Mesorhizobium loti            ---------E--I--V----- --YD-PVTEK-KKLV-Y--VKG-- ------A   G-K-QS-S----MI-Q---E---A----
Ri. prowazekii                ---------R--IY-V----- -N-T-PM-RK-QGLV-YN-VK--- ------A   DNH-YS-S----FI-Q---E---N----
C. crescentus                 ---------T----------- -TAS-PV-EK-KGMV-YE-VKGPT -----KA   H-KDYS-QEV--FI-Q---EA--AH---
He. pylori 26695              ---------EK-IYS---IM- LM-NEDKAKEAEKRL-Y--VDR - -ACAI-I   S-KVYT-QE---KI-M-L-ED--S----
He. pylori J99                ---------EK-IYS---IM- LM-NEDKAKEAEKRL-Y--VDR - -ACAI-I   S-KIYT-QE---KI-M-L-ED--S----
Camp. jejuni                  ---------EK-IYS---IM- LMINEDAAKEAKNRL-YH-TER - -ACAI-I   A-KIYT-QE---K--M-L-ED--AF---
Chl. muridarum                ---------EK--AST--F-- -K-  S--ESEIKTV-Y-VAPNSK ---VF--   ENKLYT-EE-G-QI-M---E---A----
Chl. pneumoniae J138          ---------EK--GST--F-- -KY  S--ASEIQTV-YTVTSGSK ---VF--   D-KQYT-EE-G-QI-M---E---A----
Chlam. pneumoniae AR39        ---------EK--GST--F-- -KY  S--ASEIQTV-YTVTSGSK ---VF--   D-KQYT-EE-G-QI-M---E---A----
Chlam. pneumoniae CWL029      ---------EK--GST--F-- -KY  S--ASEIQTV-YTVTSGSK ---VF--   D-KQYT-EE-G-QI-M---E---A----
Chl. trachomatis              ---------EK--AST--F-- -K-  S--ESEIKTV-Y-VAPNSK ---VFD-   EQKLYT-EE-G-QI-M---E---A----
Aqu. aeolicus                 ---R-ILD-E--VYES--F-- -K-  ---KEEAKRVSY-VVPDEK ---AFDIPNA-KLVR-EEVG-H--R-L-EA--AF---
Tre. pallidum                 --N-M----EH-IYS---F-- S--  N-LTGEAKKV-Y--V PQG D-VR---   E-KLYSTQE---FI-Q-----------
Bo. burgdorferi               --N-M----E--IYS---FM- ---  ---ASEIKMV-Y--EKGL- ---R-NISNIKKQ-S--E---AT-T---E---A----
Synechocystis PCC6803-1       ------M--G--FYSV--F-- -K-  D-ITNEATEVAYSVVKDG- -NVKLDCPAQ-KQF--EE---Q--R-LVDD-SK----
Synechocystis PCC6803-2       ----S---AE--VYS---F-- --W  DDTVEER-RV-YNCVKGRD DTVS-SI   R--SYT-QE---MI-Q-L-ADS-AF---
Deino. radiodurans            -R---AL--AA---EV--F-- --W  D--KEEAARS--TVKEGPS -SVRI--   N-KDL--E-V-----R-LVSD-SAK--N
Thermotoga maritima           ----MIL--ER-IKS---KM-                          T-YK-RI   DDKEYT-QE---FI---L-ND--A---G
Myc. tuberculosis H37Rv       --N-----VDR-VRSV--HM-                          S-WSI-I   D-K-YTA-E---RI-M-L-RD--A----
Myc. tuberculosis CDC1551     --N-----VDR-VRSV--HM-                          S-WSI-I   D-K-YTA-E---RI-M-L-RD--A----
Myc. leprae                   --N-----VDR-IRSV--HM-                          S-WSI-I   D-K-YTAQE---R--M-L-RD--A----
Bac. subtilis                 ----SI---N -IMS---HM-                          T-YK--I   E-KDYT-QEV--II-QHL-SY--S----
Bac. halodurans               -----I---N -VIS---HM-                          TNHKENI   E-KEYT-QE---II-Q-L-SD--A----
Sta. aureus MU50              -----I---N -VQS---HM-                          T-YK-DI   E-KSYT-QE---MI-QNL-N---S----
Sta. aureus N315              -----I---N -VQS---HM-                          T-YK-DI   E-KSYT-QE---MI-QNL-N---S----
L. lactis                     ---------E -IIS--SKM-                          TSEK-SA   N-KEYT-QE---MI-QNL-A---A----
Strep. pyogenes               ---------E -VIS--SKM-                          TSEK-SA   N-KEYT-QE---MI-QYL-GY-------
U. urealyticum                ---KQI---N -ISS----M-                          TKEK-T-   LNKDYT-EE---KI-TYI-EY--KKI-A
M. pulmonis                   ----LE---DT IAS----M-                          TTKTVKA   N-KTYK-EE---MI-SHL-EY--KKV-K
M. genitalium                 ----M----N -IVS----M-                          TSNK-K-§ TTKELS-E-V--QI-SYL-DF--KKI-K
M. pneumoniae                 ----M----N -IVS----M-                          TSNK-T-§ STKELT-EEV--QI-SYL-DY--KKI-K
```

Fig. 3 Alignment of Hsp70 homologues from completed bacterial genomes, showing the large insert (*boxed*) characteristic of Gram-negative bacteria

protein, MreB, which corresponds to the N-terminal half of Hsp70 [25]. Since the MreB protein, which is believed to have evolved independently from an ancestor of the Hsp70 family of proteins, does not contain this indel, the form of Hsp70 lacking the indel is indicated to be ancestral [24, 25]. Another argument in support of this view can be made on the basis of the cell structure of the prokaryotic organisms. In the formation of the ancestral prokaryotic cell, membrane enclosure must have been a key event [45]. The initial membrane enclosure probably consisted of a single unit membrane, as found in Gram-positive bacteria and archaebacteria, rather than of two different membranes separated by an intervening compartment, as found in Gram-negative bacteria [19, 22]. All of these observations indicate that the Gram-positive group lacking the large indel in Hsp70 is ancestral, in comparison with Gram-negative bacteria. The rooting based on these observations provides a useful reference point for interpreting the signature sequences in various other proteins and for deducing the relative branching orders of different groups. Based on this rooting, it could now be inferred that the 13-amino-acid indel in the S12 protein (Fig. 2), which is present in the low G + C Gram-positive bacteria (also archaebacteria) [19], but absent from both high G + C Gram-positive bacteria and different Gram-negative bacteria, is a deletion in the common ancestor of the latter groups of species. This in turn indicates that, in comparison with the high G + C group, the low G + C group is ancestral [19].

Hsp60/GroEL protein

The Hsp60/GroEL family of proteins found in all sequenced bacterial genomes contain a 1-amino-acid insert in a highly conserved region which is indicated to have been introduced after the branching of various Gram-positive bacteria and the *Deinococcus–Thermus* groups (Fig. 1) [19]. Among the completed bacterial genomes, this insert was not found in any of the Gram-positive bacterial homologues or in *D. radiodurans*, but it was present in all other bacteria (Fig. 4). Several Gram-positive bacteria contain multiple Hsp60 homologues and this insert was not present in any of them. Similarly, *Mesorhizobium loti* and other members of the Rhizobiaceae family contain multiple Hsp60 homologues and this insert is present in all of them. The indicated position of this signature is highly reliable as, of more than 300 bacterial Hsp60 sequences that are available in databases, no exceptions are observed [23].

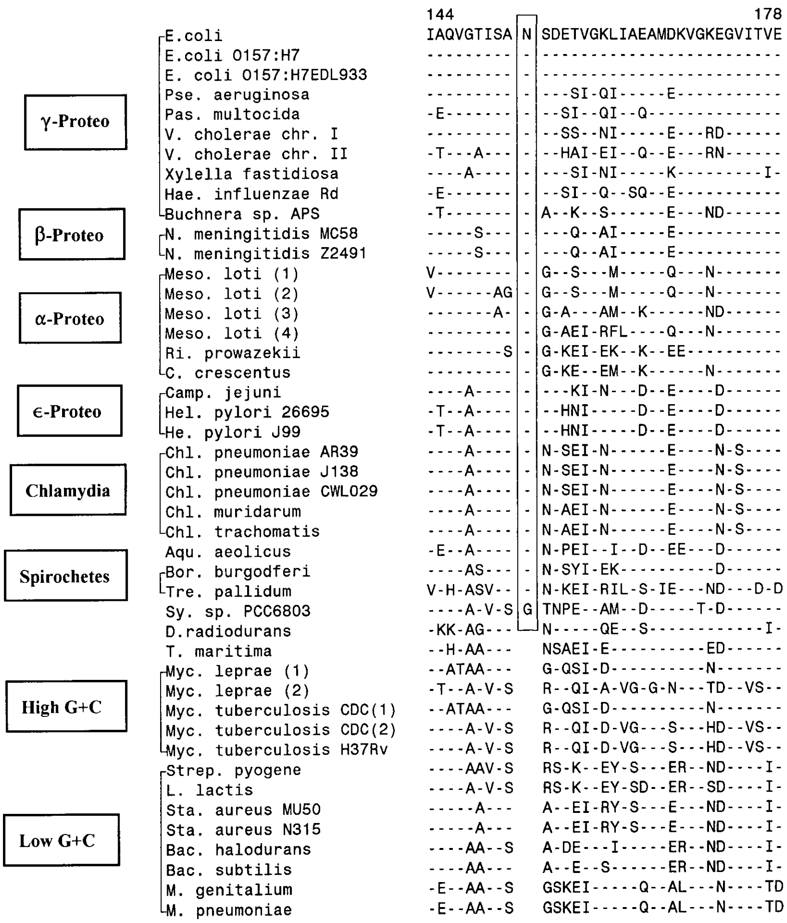


Fig. 4 Alignment of Hsp60 homologues from bacterial genomes, showing a 1-amino-acid insert (*boxed*) that was introduced after the branching of Gram-positive bacteria and the *Deinococcus–Thermus* groups

```
                                 144                                   178
E.coli                           IAQVGTISA N SDETVGKLIAEAMDKVGKEGVITVE
E.coli 0157:H7                   --------- - -------------------------
E. coli 0157:H7EDL933            --------- - -------------------------
Pse. aeruginosa                  --------- - ---SI-QI-----E-----------
Pas. multocida                   -E------- - --SI--QI--Q--------------
V. cholerae chr. I               --------- - --SS--NI-----E---RD------
V. cholerae chr. II              -T---A--- - --HAI-EI--Q--E---RN------
Xylella fastidiosa               ----A---- - ---SI-NI-----K---------I-
Hae. influenzae Rd               -E------- - --SI--Q--SQ--E-----------
Buchnera sp. APS                 -T------- - A--K--S------E---ND------
N. meningitidis MC58             -----S--- - ---Q--AI-----E-----------
N. meningitidis Z2491            -----S--- - ---Q--AI-----E-----------
Meso. loti (1)                   V-------- - G--S---M-----Q---N-------
Meso. loti (2)                   V------AG - G--S---M-----Q---N-------
Meso. loti (3)                   -------A- - G-A---AM--K------ND------
Meso. loti (4)                   --------- - G-AEI-RFL----Q---N-------
Ri. prowazekii                   --------S - G-KEI-EK--K--EE----------
C. crescentus                    --------- - G-KE--EM--K------N-------
Camp. jejuni                     ----A---- - ---KI-N---D--E----D------
Hel. pylori 26695                -T--A---- - --HNI-----D--E----D------
He. pylori J99                   -T--A---- - --HNI-----D--E----D------
Chl. pneumoniae AR39             ----A---- - N-SEI-N------E----N-S----
Chl. pneumoniae J138             ----A---- - N-SEI-N------E----N-S----
Chl. pneumoniae CWL029           ----A---- - N-SEI-N------E----N-S----
Chl. muridarum                   ----A---- - N-AEI-N------E----N-S----
Chl. trachomatis                 ----A---- - N-AEI-N------E----N-S----
Aqu. aeolicus                    -E--A---- - N-PEI--I--D--EE---D------
Bor. burgodferi                  ----AS--- - N-SYI-EK----------D------
Tre. pallidum                    V-H-ASV-- - N-KEI-RIL-S-IE---ND---D-D
Sy. sp. PCC6803                  ----A-V-S G TNPE--AM--D-----T-D------
D.radiodurans                    -KK-AG---   N-----QE--S------------I-
T. maritima                      --H-AA---   NSAEI-E----------ED------
Myc. leprae (1)                  --ATAA---   G-QSI-D----------N-------
Myc. leprae (2)                  -T--A-V-S   R--QI-A-VG-G-N---TD--VS--
Myc. tuberculosis CDC(1)         --ATAA---   G-QSI-D----------N-------
Myc. tuberculosis CDC(2)         ----A-V-S   R--QI-D-VG---S---HD--VS--
Myc. tuberculosis H37Rv          ----A-V-S   R--QI-D-VG---S---HD--VS--
Strep. pyogene                   ----AAV-S   RS-K--EY-S---ER--ND----I-
L. lactis                        ----A-V-S   RS-K--EY-SD--ER--SD----I-
Sta. aureus MU50                 -----A---   A--EI-RY-S---E---ND----I-
Sta. aureus N315                 -----A---   A--EI-RY-S---E---ND----I-
Bac. halodurans                  ----AA--S   A-DE---I-----ER--ND----I-
Bac. subtilis                    ----AA---   A--E--S------ER--ND----I-
M. genitalium                    -E--AA--S   GSKEI-----Q--AL---N----TD
M. pneumoniae                    -E--AA--S   GSKEI-----Q--AL---N----TD
```

FtsZ protein

The homologues of the FtsZ protein, which is involved in bacterial cell division, are found in all completed bacterial genomes, except those of the mycoplasma and *Chlamydiae* spp, which are intracellular pathogens [15, 17, 28, 55, 62]. A 1-amino-acid insert in a highly conserved region of this protein is indicated to have been introduced after the branching of Gram-positive bacteria, the *Deinococcus–Thermus* group, and Cyanobacteria (Fig. 1). As expected, this insert was not found in any Gram-positive bacteria, *D. radiodurans* or *Synechocystis* sp., but it was present in all other bacterial species, including *Aquifex*, Spirochetes, and different groups of proteobacteria (Fig. 5).

Alanyl-tRNA synthetase

Alanyl-tRNA synthetase contains a 4-amino-acid insert which is commonly shared by all proteobacteria and by the *Aquifex*, *Chlamydia*, and the *Cytophaga–Flavobacteria*–green sulfur bacteria groups, but is absent from all other Bacteria and Archaea (Fig. 6) [26]. This insert is indicated to have been introduced in a common ancestor of the above groups after the branching of Gram-positive bacteria, *Deioncoccus–Thermus*, Cyanobacteria, and Spirochetes (Fig. 1). Alanyl-tRNA synthetase is found in all sequenced bacterial genomes and the presence or absence of this signature in various species followed the expected pattern, with no exceptions observed (Fig. 6).

Signature sequences for proteobacteria in Hsp70 and CTP synthase

The Hsp70 protein discussed above contains a 2-amino-acid insert, within the large insert found in the Gram-negative bacteria, which is commonly shared by all proteobacteria but not found in any other bacteria [19]. In the completed bacterial genomes, this insert was present in the Hsp70 homologues from all 17 proteobacterial species, but none of the other bacteria (Fig. 7). The sequences from Gram-positive bacteria lacking this region are not shown in this figure. The enzyme CTP synthase, found in all sequenced bacterial genomes except for the mycoplasma species, contains a 10-amino-acid insert which is specific for proteobacteria (Fig. 8). This insert was found in all sequenced proteobacterial genomes but not in any other species. A smaller 4-amino-acid insert in CTP synthase that is specific for the mycobacterial species

Fig. 5 Alignment of FtsZ homologues, showing a 1-amino-acid insert (*boxed*) that was introduced after the branching of Gram-positive bacteria, the *Deinococcus–Thermus* group and the cyanobacteria

```
                                               232                      269
               E.coli K12                      GEDRAEEAAEMAISSPLLE D IDLSGARGVLVNITAGFD
               E. coli O157:H7                 ------------------- - ------------------
               E. coli O157:H7 EDL933          ------------------- - ------------------
               Vibrio cholerae                 ------------------- - ---A-----------L-
γ-Proteo       Buchnera sp. APS                --N-----S-I-------- - ------------------
               Xylella fastidiosa              -D---QA---A-VQN---D - VN-A--N-I-------S-
               Pas. multocida                  --G-----TRI-VK-D--- R V-----K-------S-M-
               H. influenzae                   --G------RL-VRND--- I K---N-Q-I-------M-
               Pse.aeruginosa                  -PN--R--T-A--RN---- - VN-Q----I-------P-
β-Proteo       N.meningitidis Z2491            -I---RM-TDQ-------D - VT-D----------TAPG
               N. meningitidis MC58            -I---RM-TDQ-------D - VT-D----------TAPG
               Ri. prowazekii                  -----IK---S---N---D H SSMC------I---G-P-
α-Proteo       Meso. loti                      --S--MK---A--AN---D E VSMK--K----S-SG-R-
               C. crescentus                   A----LM--QN--AN---D E VS-K--KA----V-G-M-
               Camp. jejuni                    --NAI---LSN--E----D G M-IK--K--ILHFKTSSN
ε-Proteo       Hel. pylori 26695               --ES-KL-VQN--Q----D - ASIE--KSII-FFEHHP-
               Hel. pylori J99                 --ES-KL-VQN--Q----D - ASIE--KSII-FFEHHP-
               Aqu. aeolicus                   -DEK-DI-V-K-VT----- G NT-E---RL--T-WTSE-
Spirochetes    Tre. pallidum                   --N--VD--TA--NN---- E TRIE--TRL--AVRGSEN
               Bor.burgdorferi                 --N--VDRRTS---N---- E VRIE-SK-L---V-G-D-
               Sy.sp.PCC6803                   -KS--K---TA--------   SSIQ--K--VF-V-G-T-
               D. radiodurans                  -DKM-----MS--H-----   RGIE---RI---VTG-Y-
               T. maritima                     --H--R---KK-ME-K-I-   HPVEN-SSIVF----PSN
               Myc. leprae                     -DG-SLK---I--N-----   ASME--Q---MS-AG-S-
High G+C       Myc.tuberculosis H37 Rv         --G-SLK---I--N-----   ASME--Q---MS-AG-S-
               Myc.tuberculosis CDC1551        --G-SLK---I--N-----   ASME--Q---MS-AG-S-
               Sta.aureus Mu50                 --N--V---KK--------   TSIV--Q---M---G-ES
               Sta. aureus Z2491               --N--V---KK--------   TSIV--Q---M---G-ES
Low G+C        Bac.subtilis                    --N--A---KK--------   AAID--Q---M---G-TN
               Bac. halodurans                 --N--G---KK--------   TS-D--Q---M---G-SN
               M. pulmonis                     -K---VK--IH-----II-   TSIQ--SHTII---GSAN
               Lac. lactis                     --E-VI--TRK--Y-----   TTIE--EN--L-V-G-M-
               Strep. pyogenes                 --E-IV---RK--Y-----   TTID--QD-I--V-G-L-
```

is found in the same position as the proteobacterial insert (Fig. 8). However, this insert, because of its size and specificity, is of independent origin and it does not confuse or affect the specificity of the proteobacterial signature.

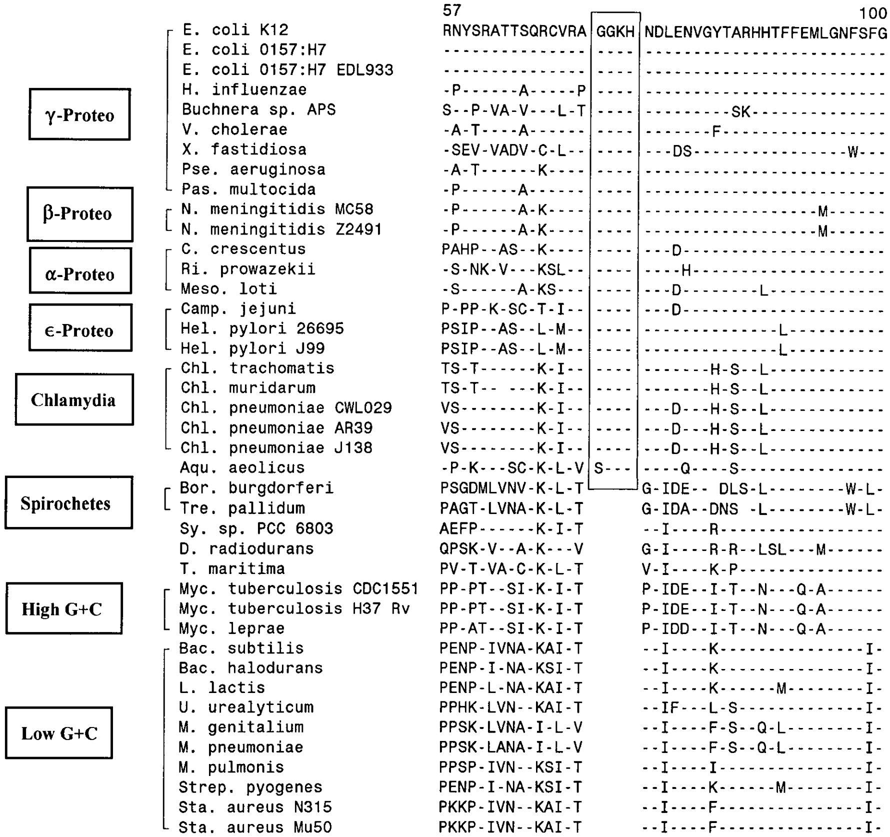
Signature sequences indicating the branch order of the proteobacterial groups

Signature sequences in a number of proteins have been shown to make clear distinctions among different groups of proteobacteria [23]. A 1-amino-acid conserved insert in the Lon protease is commonly shared by all α-, β-, and γ-proteobacterial species but not present in any other species. Lon protease homologues are present in all bacterial genomes, except a few Gram-positive bacteria. The insert in Lon protease, as expected, was found in all α-, β-, and γ-proteobacterial species but not in any other species (Fig. 9). Another signature introduced at a similar stage is found in the SecA protein. The SecA homologues are found in all sequenced bacterial genomes and the 7-amino-acid insert is seen in all of the α-, β-, and γ-proteobacteria but not in any other bacteria (Fig. 10). A smaller insert in this position is also seen in the two spirochete species but, based on its size and species specificity, this insert was probably introduced independently. The genomes from chlamydial species contain another SecA related protein (not shown), which contains a very large insert in this region, quite different from the insert found in α-, β-, and γ-proteobacteria.

The Hsp70 family of proteins contains another useful signature that is distinctive of the β- and γ-proteobacteria. This signature, consisting of a 4-amino-acid insert in a highly conserved region, is found in all of the β- and γ-proteobacterial species from sequenced genomes but not in any other species (Fig. 11). The β- and γ-proteobacterial species, in addition to the orthologous Hsp70 protein, also contain a protein, Hsc66, which is distantly related to Hsp70 and carries out unrelated functions [34, 57]. The Hsc66 homologues, do not contain the β- or γ-insert, but they are readily distinguished from the Hsp70 homologues because of extensive sequence divergence in different regions, particularly towards the C-terminal end. Another signature, a 1-amino-acid insert, distinctive of the β- and γ-proteobacteria, has been identified in the protein, phosphoribosyl pyrophosphate synthetase. Among the sequenced bacterial genomes, this signature is found in all β- and γ-proteobacteria but in none of the other species (Fig. 12). The γ-proteobacterial group differs from other proteobacteria by a 2-amino-acid deletion in the enzyme, 5′-phosphoribosyl-5-aminoimidazol-4-carboxamide transformylase. This deletion was found in all of the γ-proteobacterial genomes (Fig. 13), but in none of the other species where the homologues of this protein are found. In *T. maritima*, a large deletion of 12–13 amino acids is present in this position which probably originated independently.

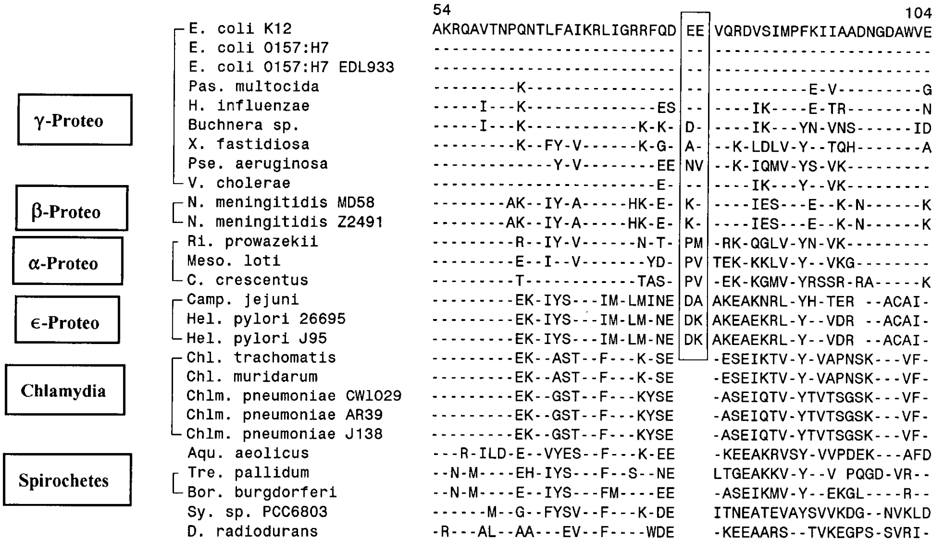
The distribution of indels in genomic sequences strongly supports the indel model

The question could now be asked whether the observed results from genomic sequences support the evolutionary model based on indels, or whether these results can be explained by any other reasonable mechanism. In the evolutionary model based on indels, there are two potential problems that could give misleading results. First, it is possible that a given indel, rather than being derived from a common ancestor, was introduced on



```
                             57                                      100
E. coli K12                  RNYSRATTSQRCVRA GGKH NDLENVGYTARHHTFFEMLGNFSFG
E. coli 0157:H7              --------------- ---- -------------------------
E. coli 0157:H7 EDL933       --------------- ---- -------------------------
H. influenzae                -P------A-----P ---- -------------------------
Buchnera sp. APS             S--P-VA-V---L-T ---- ---------SK--------------
V. cholerae                  -A-T----A------ ---- -------F-----------------
X. fastidiosa                -SEV-VADV-C-L-- ---- ---DS----------------W---
Pse. aeruginosa              -A-T------K---- ---- -------------------------
Pas. multocida               -P------A------ ---- -------------------------
N. meningitidis MC58         -P------A-K---- ---- ------------------M------
N. meningitidis Z2491        -P------A-K---- ---- ------------------M------
C. crescentus                PAHP--AS--K---- ---- ---D---------------------
Ri. prowazekii               -S-NK-V---KSL-- ---- ----H--------------------
Meso. loti                   -S------A-KS--- ---- ---D--------L------------
Camp. jejuni                 P-PP-K-SC-T-I-- ---- ---D---------------------
Hel. pylori 26695            PSIP--AS--L-M-- ---- --------------L----------
Hel. pylori J99              PSIP--AS--L-M-- ---- --------------L----------
Chl. trachomatis             TS-T------K-I-- ---- -------H-S--L------------
Chl. muridarum               TS-T-- ---K-I-- ---- -------H-S--L------------
Chl. pneumoniae CWL029       VS--------K-I-- ---- ---D---H-S--L------------
Chl. pneumoniae AR39         VS--------K-I-- ---- ---D---H-S--L------------
Chl. pneumoniae J138         VS--------K-I-- ---- ---D---H-S--L------------
Aqu. aeolicus                -P-K---SC-K-L-V S--- ----Q----S---------------
Bor. burgdorferi             PSGDMLVNV-K-L-T      G-IDE-- DLS-L--------W-L-
Tre. pallidum                PAGT-LVNA-K-L-T      G-IDA--DNS -L--------W-L-
Sy. sp. PCC 6803             AEFP------K-I-T      --I----R-----------------
D. radiodurans               QPSK-V--A-K---V      G-I----R-R--LSL---M------
T. maritima                  PV-T-VA-C-K-L-T      V-I----K-P---------------
Myc. tuberculosis CDC1551    PP-PT--SI-K-I-T      P-IDE--I-T--N---Q-A------
Myc. tuberculosis H37 Rv     PP-PT--SI-K-I-T      P-IDE--I-T--N---Q-A------
Myc. leprae                  PP-AT--SI-K-I-T      P-IDD--I-T--N---Q-A------
Bac. subtilis                PENP-IVNA-KAI-T      --I----K---------------I-
Bac. halodurans              PENP-I-NA-KSI-T      --I----K---------------I-
L. lactis                    PENP-L-NA-KAI-T      --I----K------M--------I-
U. urealyticum               PPHK-LVN--KAI-T      --IF---L-S-------------I-
M. genitalium                PPSK-LVNA-I-L-V      --I----F-S--Q-L--------I-
M. pneumoniae                PPSK-LANA-I-L-V      --I----F-S--Q-L--------I-
M. pulmonis                  PPSP-IVN--KSI-T      --I----I---------------I-
Strep. pyogenes              PENP-I-NA-KSI-T      --I----K------M--------I-
Sta. aureus N315             PKKP-IVN--KAI-T      --I----F---------------I-
Sta. aureus Mu50             PKKP-IVN--KAI-T      --I----F---------------I-
```

Fig. 6 Alignment of Ala-tRNA synthetase sequences, showing a 4-amino-acid insert (*boxed*) that is common to only the *Chlamydiae–Aquifex* group and proteobacterial species and is not found in any other groups of bacteria

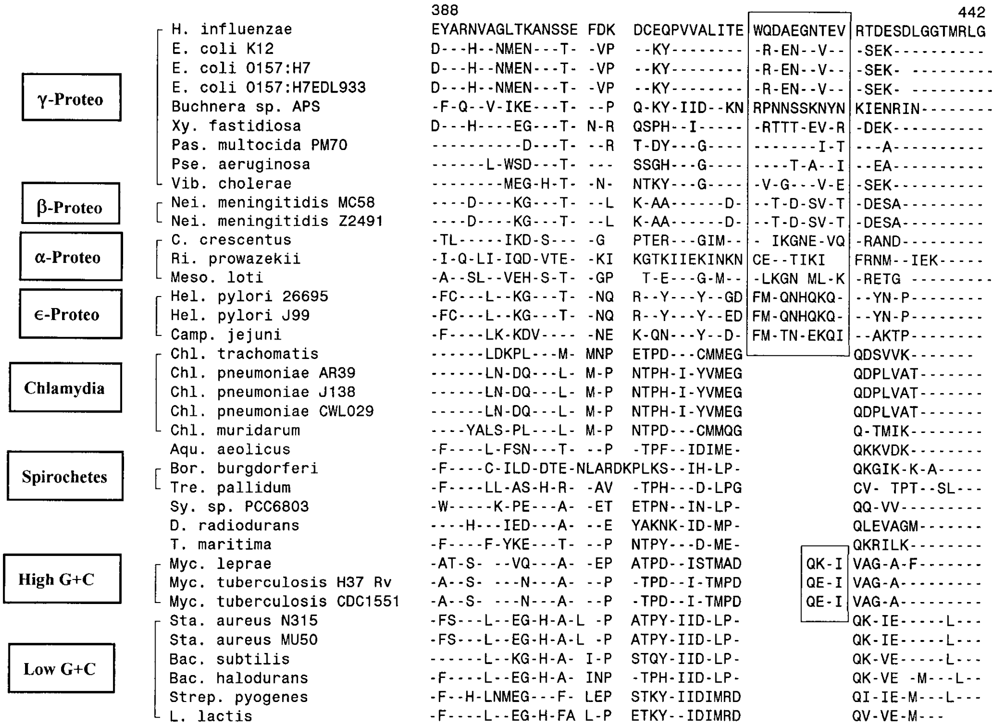


```
                            54                                                   104
E. coli K12                 AKRQAVTNPQNTLFAIKRLIGRRFQD EE VQRDVSIMPFKIIAADNGDAWVE
E. coli 0157:H7             -------------------------- -- -----------------------
E. coli 0157:H7 EDL933      -------------------------- -- -----------------------
Pas. multocida              ---------K---------------- -- ----------E-V---------G
H. influenzae               -----I---K--------------ES -- ----IK----E-TR--------N
Buchnera sp.                -----I---K------------K-K- D- ----IK---YN-VNS------ID
X. fastidiosa               ---------K--FY-V------K-G- A- --K-LDLV-Y--TQH-------A
Pse. aeruginosa             -------------Y-V--------EE NV --K-IQMV-YS-VK---------
V. cholerae                 ------------------------E- -- ----IK---Y--VK---------
N. meningitidis MD58        --------AK--IY-A-----HK-E- K- ----IES---E--K-N------K
N. meningitidis Z2491       --------AK--IY-A-----HK-E- K- ----IES---E--K-N------K
Ri. prowazekii              ---------R--IY-V------N-T- PM -RK-QGLV-YN-VK---------
Meso. loti                  ---------E--I--V-------YD- PV TEK-KKLV-Y--VKG--------
C. crescentus               ---------T------------TAS- PV -EK-KGMV-YRSSR-RA-----K
Camp. jejuni                ---------EK-IYS---IM-LMINE DA AKEAKNRL-YH-TER --ACAI-
Hel. pylori 26695           ---------EK-IYS---IM-LM-NE DK AKEAEKRL-Y--VDR --ACAI-
Hel. pylori J95             ---------EK-IYS---IM-LM-NE DK AKEAEKRL-Y--VDR --ACAI-
Chl. trachomatis            ---------EK--AST--F---K-SE    -ESEIKTV-Y-VAPNSK---VF-
Chl. muridarum              ---------EK--AST--F---K-SE    -ESEIKTV-Y-VAPNSK---VF-
Chlm. pneumoniae CW1029     ---------EK--GST--F---KYSE    -ASEIQTV-YTVTSGSK---VF-
Chlm. pneumoniae AR39       ---------EK--GST--F---KYSE    -ASEIQTV-YTVTSGSK---VF-
Chlm. pneumoniae J138       ---------EK--GST--F---KYSE    -ASEIQTV-YTVTSGSK---VF-
Aqu. aeolicus               ---R-ILD-E--VYES--F---K-EE    -KEEAKRVSY-VVPDEK---AFD
Tre. pallidum               --N-M----EH-IYS---F--S--NE    LTGEAKKV-Y--V PQGD-VR--
Bor. burgdorferi            --N-M----E--IYS---FM----EE    -ASEIKMV-Y--EKGL----R--
Sy. sp. PCC6803             ------M--G--FYSV--F---K-DE    ITNEATEVAYSVVKDG--NVKLD
D. radiodurans              -R---AL--AA---EV--F----WDE    -KEEAARS--TVKEGPS-SVRI-
```

Fig. 7 Alignment of Hsp70 homologues from bacterial genomes, showing a 2-amino-acid insert (*boxed*) that is commonly found in all proteobacterial species. The Hsp70 homologues from Gram-positive bacteria lack this region and hence are not shown

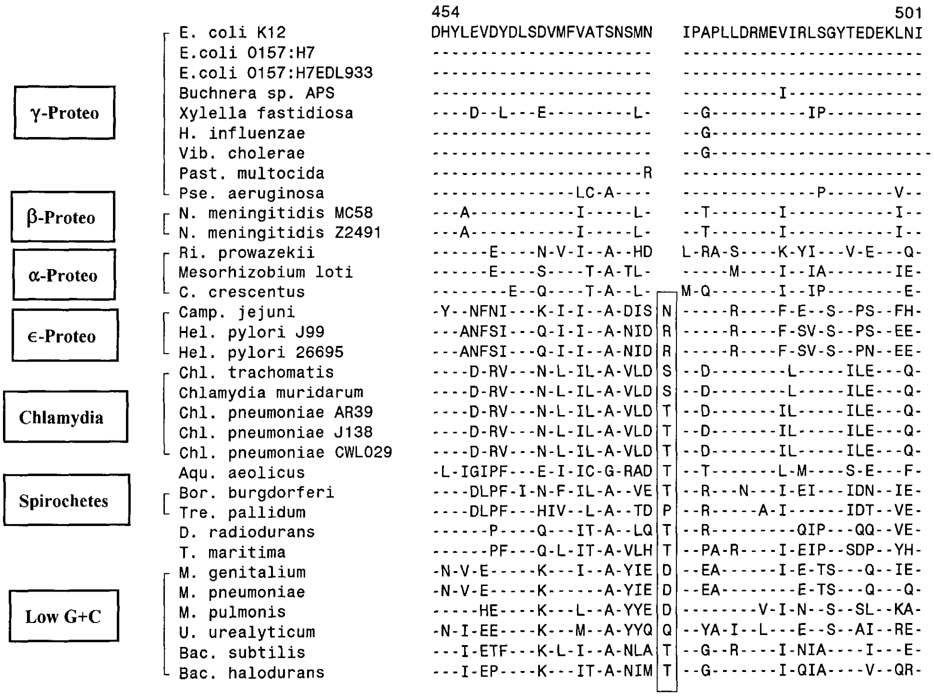
multiple occasions in different species/groups due to similar functional constraints operating on the protein. Second, the shared presence of an indel in different species could also occur if the indel was originally introduced in one species (or group of species) and then transferred to others by LGT. The analyses of genomic sequences in the past few years have led to the view that LGT among prokaryotic species is quite common and that it poses a major problem in deducing evolutionary relationships among prokaryotes [3, 11, 30, 39, 71].

The basic premise on which the indel model is based is that, once an indel has been introduced in an ancestral lineage, various species emerging from that ancestor henceforth should all contain the indel, whereas all species from different groups that either existed prior to the introduction of the indel or which did not evolve from this ancestor should lack the indel. In contrast, if these indels have been introduced into various groups independently or if the genes containing these indels have undergone frequent LGT from one species to



```
                                  388                                                             442
H. influenzae                     EYARNVAGLTKANSSE FDK  DCEQPVVALITE WQDAEGNTEV RTDESDLGGTMRLG
E. coli K12                       D---H--NMEN---T- -VP  --KY-------- -R-EN--V-- -SEK----------
E. coli O157:H7                   D---H--NMEN---T- -VP  --KY-------- -R-EN--V-- -SEK- ---------
E. coli O157:H7EDL933             D---H--NMEN---T- -VP  --KY-------- -R-EN--V-- -SEK- ---------
Buchnera sp. APS                  -F-Q--V-IKE---T- --P  Q-KY-IID--KN RPNNSSKNYN KIENRIN-------
Xy. fastidiosa                    D---H----EG---T- N-R  QSPH--I----- -RTTT-EV-R -DEK----------
Pas. multocida PM70               ----------D---T- --R  T-DY---G---- -------I-T ---A----------
Pse. aeruginosa                   ------L-WSD---T- ---  SSGH---G---- ----T-A--I --EA----------
Vib. cholerae                     --------MEG-H-T- -N-  NTKY---G---- -V-G---V-E -SEK----------
Nei. meningitidis MC58            ----D----KG---T- --L  K-AA------D- --T-D-SV-T -DESA---------
Nei. meningitidis Z2491           ----D----KG---T- --L  K-AA------D- --T-D-SV-T -DESA---------
C. crescentus                     -TL-----IKD-S--- -G   PTER---GIM-- - IKGNE-VQ -RAND---------
Ri. prowazekii                    -I-Q-LI-IQD-VTE- -KI  KGTKIIEKINKN CE--TIKI   FRNM--IEK-----
Meso. loti                        -A--SL--VEH-S-T- -GP   T-E---G-M-- -LKGN ML-K -RETG ---------
Hel. pylori 26695                 -FC---L--KG---T- -NQ  R--Y---Y--GD FM-QNHQKQ- --YN-P--------
Hel. pylori J99                   -FC---L--KG---T- -NQ  R--Y---Y--ED FM-QNHQKQ- --YN-P--------
Camp. jejuni                      -F----LK-KDV---- -NE  K-QN---Y--D- FM-TN-EKQI --AKTP--------
Chl. trachomatis                  ------LDKPL---M- MNP  ETPD---CMMEG            QDSVVK-------
Chl. pneumoniae AR39              ------LN-DQ---L- M-P  NTPH-I-YVMEG            QDPLVAT-------
Chl. pneumoniae J138              ------LN-DQ---L- M-P  NTPH-I-YVMEG            QDPLVAT-------
Chl. pneumoniae CWL029            ------LN-DQ---L- M-P  NTPH-I-YVMEG            QDPLVAT-------
Chl. muridarum                    ----YALS-PL---L- M-P  NTPD---CMMQG            Q-TMIK--------
Aqu. aeolicus                     -F----L-FSN---T- --P  -TPF--IDIME-            QKKVDK--------
Bor. burgdorferi                  -F----C-ILD-DTE-NLARDKPLKS--IH-LP-            QKGIK-K-A-----
Tre. pallidum                     -F----LL-AS-H-R- -AV  -TPH---D-LPG            CV- TPT--SL---
Sy. sp. PCC6803                   -W-----K-PE---A- -ET  ETPN--IN-LP-            QQ-VV---------
D. radiodurans                    ----H---IED---A- --E  YAKNK-ID-MP-            QLEVAGM-------
T. maritima                       -F----F-YKE---T- --P  NTPY---D-ME-            QKRILK--------
Myc. leprae                       -AT-S- --VQ---A- -EP  ATPD--ISTMAD       QK-I VAG-A-F-------
Myc. tuberculosis H37 Rv          -A--S- ---N---A- --P  -TPD--I-TMPD       QE-I VAG-A---------
Myc. tuberculosis CDC1551         -A--S- ---N---A- --P  -TPD--I-TMPD       QE-I VAG-A---------
Sta. aureus N315                  -FS---L--EG-H-A-L -P  ATPY-IID-LP-            QK-IE-----L---
Sta. aureus MU50                  -FS---L--EG-H-A-L -P  ATPY-IID-LP-            QK-IE-----L---
Bac. subtilis                     ------L--KG-H-A- I-P  STQY-IID-LP-            QK-VE-----L---
Bac. halodurans                   -F----L--EG-H-A- INP  -TPH-IID-LP-            QK-VE -M---L--
Strep. pyogenes                   -F--H-LNMEG---F- LEP  STKY-IIDIMRD            QI-IE-M---L---
L. lactis                         -F----L--EG-H-FA L-P  ETKY--IDIMRD            QV-VE-M---
```

Fig. 8 Sequence alignment of CTP synthetase from bacterial genomes, showing a 10-amino-acid insert (*boxed*) common to all proteobacterial groups



```
                             454                                                   501
E. coli K12                  DHYLEVDYDLSDVMFVATSNSMN    IPAPLLDRMEVIRLSGYTEDEKLNI
E.coli O157:H7               -----------------------    -------------------------
E.coli O157:H7EDL933         -----------------------    -------------------------
Buchnera sp. APS             -----------------------    ----------I--------------
Xylella fastidiosa           ----D--L---E---------L-    --G----------IP----------
H. influenzae                -----------------------    --G----------------------
Vib. cholerae                -----------------------    --G-----------------------
Past. multocida              ----------------------R    -------------------------
Pse. aeruginosa              ---------------LC-A----    --------------P-------V--
N. meningitidis MC58         ---A-----------I-----L-    --T-------I-----------I--
N. meningitidis Z2491        ---A-----------I-----L-    --T-------I-----------I--
Ri. prowazekii               ------E----N-V-I--A--HD    L-RA-S----K-YI---V-E---Q-
Mesorhizobium loti           ------E----S----T-A-TL-    -----M----I--IA-------IE-
C. crescentus                --------E--Q----T-A--L-    M-Q-------I--IP--------E-
Camp. jejuni                 -Y--NFNI---K-I-I--A-DIS N  -----R----F-E--S--PS--FH-
Hel. pylori J99              ---ANFSI---Q-I-I--A-NID R  -----R----F-SV-S--PS--EE-
Hel. pylori 26695            ---ANFSI---Q-I-I--A-NID R  -----R----F-SV-S--PN--EE-
Chl. trachomatis             ----D-RV---N-L-IL-A-VLD S  --D--------L-----ILE---Q-
Chlamydia muridarum          ----D-RV---N-L-IL-A-VLD S  --D--------L-----ILE---Q-
Chl. pneumoniae AR39         ----D-RV---N-L-IL-A-VLD T  --D-------IL-----ILE---Q-
Chl. pneumoniae J138         ----D-RV---N-L-IL-A-VLD T  --D-------IL-----ILE---Q-
Chl. pneumoniae CWL029       ----D-RV---N-L-IL-A-VLD T  --D-------IL-----ILE---Q-
Aqu. aeolicus                -L-IGIPF---E-I-IC-G-RAD T  --T-------L-M----S-E---F-
Bor. burgdorferi             ----DLPF-I-N-F-IL-A--VE T  --R---N---I-EI---IDN--IE-
Tre. pallidum                ----DLPF---HIV--L-A--TD P  --R-----A-I------IDT--VE-
D. radiodurans               ------P----Q---IT-A--LQ T  --R---------QIP---QQ--VE-
T. maritima                  ------PF---Q-L-IT-A-VLH T  --PA-R----I-EIP--SDP--YH-
M. genitalium                -N-V-E-----K---I--A-YIE D  --EA------I-E-TS---Q--IE-
M. pneumoniae                -N-V-E-----K------A-YIE D  --EA--------E-TS---Q---Q-
M. pulmonis                  -----HE----K---L--A-YYE D  --------V-I-N--S--SL--KA-
U. urealyticum               -N-I-EE----K---M--A-YYQ Q  --YA-I--L---E--S--AI--RE-
Bac. subtilis                ---I-ETF---K-L-I--A-NLA T  --G--R----I-NIA----I---E-
Bac. halodurans              ---I-EP----K---IT-A-NIM T  --G-------I-QIA----V--QR-
```

Fig. 9 Alignment of Lon protease sequences from bacterial genomes, showing a 1-amino-acid insert (*boxed*) that is commonly shared by all α-, β-, and γ-proteobacteria

another, then the presence or absence of these indels in different species will not follow any predicted pattern. In such a case, different groups of species, or even individual species from different groups, will either contain or lack the indels.

A summary of the results for the various indels studied in this work is presented in Table 2. For each of the proteins containing these indels, the number of species where the protein was found is indicated, together with the number of species in which the indel was expected to be present or absent according to the model. The last column indicates the number of exceptions observed where the presence or absence of an indel was not in accordance with the indel model. As seen from Table 2, the proteins containing these indels are widely represented in different bacteria and many of them were found in all sequenced bacterial genomes. A few of these proteins are absent from species such as

		263		328
γ-Proteo	E. coli K12	FSVDEKSRQVNLTERGLVLIEELL	VKEGIMD	EGESLYSPANIMLMHHVTAALRAHALFTRDVDYIV
	E.coli O157:H7	------------------------	-------	-----------------------------------
	E.coli O157:H7EDL933	------------------------	-------	-----------------------------------
	Buchnera sp. APS	--I----K-IY------IKV-KI-	FDKKL-N	T------SN--I-----LS-----K--V-N---L-
	Pas. multocida	-TL-L-TK-AH-----QEKV-QW-	TEQ-L-S	AE------SK-S-L---Y------T--E-------
	Xylella fastidiosa	----GK--H-S-V-MERA----	HQA--LG	-ED---AAQ-LSVV--LN--------YQ-------
	H. influenzae	-TL-L--K-AH-----QEKV-DW-	IAQ-L-P	--D-----SR-V-L---M------T--EK------
	Vib. cholerae	-T-----K--H---T-QEFV----	--N-M-Q	--DT-------S-L---N------V--EKN-----
	Pse. aeruginosa	Y-I---T---E-N-Q-HQF--D--	SQN-LLG	-------AH-LS-LT--Y------T--H-N-E---
β-Proteo	N. meningitidis MC58	YW----AH--I-S-A-HEHA-QI-	TQM-LLA	-ND----A---A----LM------T--HK-QH-VI
	N. meningitidis Z2491	YW----AH--I-S-A-HEHA-QI-	TQM-LLA	-ND----A---A----LM------T--HK-QH-VI
α-Proteo	Ri. prowazekii	-EK---LKTI----A-ITH--S--	I-D--IK	PDTG--DFE-LN-V-Y-NQ-----HM--I----L-
	Mesorhizobium loti	YEI---QKTSIF--E-TEKL-N--	RDADLL	K-----DVE-VAIV---NN--K--R--QK-K----
	C. crescentus	-DH---QK--I---D-QEK---I-	MSANLAE	DSAG--RR--VSVV---NQ----NI-I---K----
ϵ-Proteo	Camp. jejuni	-V----N-NILI--A-IAKA-K-F		GV-N---LD-AI-A-QLDQ--K--N--EK--H-VL
	Hel. pylori, J99	-TI---N-AILI--E-IKKA-N-F		GVDN--KIE-AA-S--LDQ--K-NY--FI-K----
	Hel. pylori 26695	-TI---N-AILI--E-IKKA-N-F		GVDN--KIE-AA-S--LDQ--K-NY--FI-K----
Chlamydia	Chl.Pneumoniae AR39	RGL-SFLDVDI-PKDKK--EGISE		FCR--WLVSKG-PLNR-LRRV-E-PDLRAMI-KWD
	Chl.Pneumoniae J138	RGL-SFLDVDI-PKDKK--EGISE		FCR--WLVSKG-PLNR-LRRV-E-PDLRAMI-KWD
	Chl.Pneumoniae CWL029	RGL-SFLDVDI-PKDKK--EGISE		FCR--WLVSKG-PLNR-LRRV-E-PDLRAMI-KWD
	Chl.trachomatis	RGL-SFLDVDI-PKDKK--EGISE		FCR--WLVSKG-PLNR-LRRV-E-PDLRAMI-KWD
	Chl. muridarum	KTL-PFLGMDV-PKDRKIMEGISE		ACRA-WLVSKG-PLNR-LRRV-E-PDLRAM--KWD
	Aqu. aeolicus	-T----N-T-----Q-IKKV-KM-		GIDN--DLKHVD-L-AILQSI---H--KK--H---
Spirochetes	Tre. pallidum	YI----N-K-SFSGP-MLH-QDV-	THAGL	IQG--FDEE-FKYI-YF-Q-----L-YRA----V-
	Bor. burgdorferi	YT----AKRISF-AK--NNL-Q--	VSKGI	ISG-M-TDS-FNYV-YM-Q--K--L--LKNRE---
	Sy. sp. PCC6803	YE----G-N-L--DQ-FINA-Q--		GVSD-FDSNDPWA -YIFN-IK-KE--IK--N---
	D. radiodurans	YTIE---KA-H---Q-ITK--R--		SLKD----E-MDKA-MI-Q-I--RE-YH-EK----
	T. maritima	-T----A-TII---E-VAKA-KII		GV-N--D-G-VS-LY-LIN--K-LH--KK----V-
High G+C	Myc. tuberculosis H37Rv	YE--LRK-T-GVH-K-VEFV-DQ-		GIDN--ET--SP-VSYLNN--K-KE--S--K----
	Myc. tuberculosis CDC1551	YE--LRK-T-GVH-K-VEFV-DQ-		GIDN--ET--SP-VSYLNN--K-KE--S--K----
	Myc. leprae	YDT-SD--N-H--DV-ARKV-KA-		G-ID---EEHVGTTLTEVN--H--V-LQ---H---
Low G+C	Sta. aureus N315	YKY---TKA-H---Q-ADKA-RMF		KV-N--DVQ-VDVIS-INT-----VTLQ-----M-
	Sta. aureus MU50	YKY---TKA-H---Q-ADKA-RMF		KV-N--DVQ-VDVIS-INT-----VTLQ-----M-
	Bac. halodurans	YTL---TKS-Q---E-VNKA-RAF		NIDN-FDQRHVQ-L--INQ-MK--VVMH--A--V-
	Bac. subtilis	YTY-I-TKA-Q---E-MTKA-KAF		GIDN-FDVKHVA-N--INQ--K--VAMQK----V-
	Lactococcus lactis	YKI-LQ-KTIS---E-IDKA-KFF		QI-N--DME-VA-T-F-DN----NFIMLH-I--M-
	U. urealyticum	-TL-PE-QS-A---S-VEKAQKFF		NTKNY-NFE-SDII-K--N----NFT-FNGRE---
	Strep. pyogenes	YVI-VPTKTIG-SDS-IDKA-SYF		NLSN--DIE-VA-T-FIDN----NYIMLL-I--V-
	M. pneumoniae	YKI-PE--APA---L-IKHA-KNF		KTDN-FALE-SD-F-KIIN--T-VKV-EQGKE---
	M. pulmonis	YEI--E-KTIK-VDS-IDKANKFF		TLSN--DIK-SE-V-RIQN----NFIMKK--E---
	M. genitalium	-KI-PE--AAS---L-IKKA-QTF		KK-N-FALE-SD-F-KIMNG-T-VKV-EQGKE---

Fig. 10 Alignment of SecA homologues from bacterial genomes, showing a 7-amino-acid insert (*boxed*) that is common to all α-, β-, and γ-proteobacteria

mycoplasms or chlamydiae, which are intracellular pathogens, where the corresponding genes may have been lost because the cellular functions of these proteins are likely provided for by the host [15, 62]. For all of these proteins, the distribution of indels in various proteins was found to be exactly as predicted by the model, with no exceptions observed. Of a total of 450 indels whose distribution in different species was examined in the present work, all of them showed the expected distribution, as predicted by the model. The only possible exception was the presence of a second Hsp70 homologue in *B. burgdorferi*, which lacked the large insert in the protein, distinctive of Gram-negative bacteria. The BLAST searches indicate that this gene is likely acquired from Gram-positive bacteria by means of LGT. However, *B. burgdorferi* contains another Hsp70 homologue with the expected characteristics. Hence, the presence of this laterally transferred gene, which is readily identified as such and which is absent from other spirochete species, does not in any way confuse or affect the inference concerning the phylogenetic placement of this species. In a few cases, some species were found to contain a different kind of indel (differing in length, amino acid composition, species specificity) in a similar position as the indicated signature. Such indels, which are probably of independent origin, again do not confuse or affect the inference from specific indels. For all of the studied proteins, in addition to the data from completed bacterial genomes, sequence information is available from a large number of other species and, in almost all cases, the distribution of these indels in various species follows the pattern as predicted by the indel model [19, 23]. These results provide strong evidence that the inferences derived from indel data are reliable [48] and they are not affected to any significant extent by other factors, such as LGT [63] or independent occurrence of these indels in different species.

The evolutionary relationship that emerges based on indels, in addition to its high degree of internal consistency in the placement of species into different groups and in determining their relative branching order, is also quite appealing from other perspectives:

1. The model is consistent with and accounts for the major ultrastructural differences seen among the Bacteria. The model indicates 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 membrane and containing a periplasmic compartment (i.e., all true Gram-negative bacteria or diderm bacteria) [19, 47]. Of these two structurally and phylogenetically distinct groups of bacteria, the monoderm bacteria are indicated to be ancestral.
2. The model places *Deinococcus–Thermus* in an intermediate position between monoderm and diderm bacteria. This placement is consistent with the observation that *Deinococcus* contains a thick

Fig. 11 Sequence alignment of Hsp70 homologues showing a 4-amino-acid insert (*boxed*) that is distinctive of *β*- and *γ*-proteobacteria

		190		233
γ-Proteo	E. coli K12	IAVYDLGGGTFDISIIEI	DEVD	GEKTFEVLATNGDTHLGGEDFD
	E.coli 0157:H7	------------------	----	----------------------
	E.coli 0157:H7 EDL933	------------------	----	----------------------
	Pasteurella multocida	------------L-----	---G	--------S-------------
	Xylella fastidiosa	------------V-----	A---	---Q----------F-------
	Vib. cholerae	------------------	---E	--------S-------------
	Buchnera. sp. APS	------------------	----	K---------------------
	Pse. aeruginosa	VI----------V-V---	A---	--HQ----------F-------
	H. influenzae	------------------	-NF-	--Q-------G-N---------
β-Proteo	Nei. meningitidis MC58	V-----------------	ANL-	-D-Q----------F-------
	Nei. meningitidis Z2491	V-----------------	ANL-	-D-Q----------F-------
α-Proteo	Ri. prowazekii	------------V--L--		SDGV---KS-----F-------
	C. crescentus	------------V--L--		-DGV---KS-----F-------
	Mesorhizobium loti	--------------VL--		-DGV---KS-----F-------
ϵ-Proteo	Camp. jejuni	-V----------VTVL-T		-DNVV-----G-NAF---D---
	Hel. pylori 26695	-M----------VTVL-T		-DNVV-----G--AF---D---
	Hel. pylori J99	-M----------VTVL-T		-DNVV-----G--AF---D---
Chlamydia	Chl. trachomatis	---F-----------L--		-DGV----S---------D---
	Chlamydia muridarum	---F-----------L--		-DGV----S---------D---
	Chlm. pneumoniae CWL029	---F-----------L--		-DGV----S-----L---D---
	Chlm. pneumoniae AR39	---F-----------L--		-DGV----S-----L---D---
	Chlm. pneumoniae J138	---F-----------L--		-DGV----S-----L---D---
	Aqu. aeolicus	-L---F------GVS-LE		--GVI--KV-A-------ANI-
Spirochetes	Tre. pallidum	---------------L-L		-DGV---KS---------D---
	Bor. burgdorferi	V--------------L-L		-DGV---KS---------DN--
	Sy. sp. PCC6803 (1)	-L-F--------V--L-V		--GV------S-------D---
	Sy. sp. PCC6803 (2)	-L-F--------V-LLQL		-NGV----S-S-NN----D---
	D. radiodurans	VL-F--------VT-L-L		-DGV---KS-A-------A---
	T. maritima	VL----------V--L--		--GVI--I--A-NN----D---
High G+C	Myc. tuberculosis CDC1551	-L-F--------V-LL--		--GVV--R--S--N----D-W-
	Myc. tuberculosis H37 Rv	-L-F--------V-LL--		--GVV--R--S--N----D-W-
	Myc. leprae	-L-F--------V-LL--		--GVV--R--S--N----D-W-
Low G+C	Bac. halodurans	-L----------V--L-L		-DGF---K--S--NK---D---
	Bac. subtilis	-L----------V--L-L		-DGV---RS-A--NR---D---
	Strep. pyogenes	-L-F--------V--L-L		-DGV-D----A--NK---D---
	Sta. aureus MU50	VL-F--------V--L-L		-DGV----S-A--NK---D---
	Sta. aureus Z2491	VL-F--------V--L-L		-DGV----S-A--NK---D---
	L. lactis	-L-F--------V--L-L		-DGV-D----A-NNK---D---
	U. urealyticum	-L-F--------V-VLDM		ADG-----S-S--N----D-W-
	M. genitalium	VL----------V-LLD-		A-G-------A--NR---D-W-
	M. pulmonis	VL----------V-VL-L		ENG-----S-S--N----D-W-
	M. pneumoniae	VL----------V-LLD-		A-G-------A--NR---D-W-

peptidoglycan layer characteristic of Gram-positive bacteria and shows a positive Gram-staining reaction [46]. However, this species contains both inner and outer membranes, which is the main defining characteristic of Gram-negative bacteria. Thus, *Deinococcus* is indicated to be an intermediate in the transition between monoderm and diderm bacteria and it provides suggestive evidence that, in the development of Gram-negative bacteria from Gram-positive bacteria, the outer membrane evolved first, before the changes in the cell wall occurred [19].

3. For 39 of the 41 bacterial species whose genomes have been sequenced, their placement into different groups based on indel data is in agreement with that based on the 16 S rRNA. The two species (i.e., *Aquifex aeolicus*, *T. maritima*) whose phylogenetic placements differed somewhat from that based on rRNA, show very deep branching in the rRNA trees [40, 51]. Indel data places *Aquifex* in a similar position as the *Chlamydia* and *Cytophaga–Bacteriodes* groups. This inference is based on a number of different signatures, all of which place it in the same position. It is difficult to account for these results by LGT from other species [3]. The branching of *Aquifex* in a similar position as *Chalmydia* is also observed in phylogenetic trees based on a number of different proteins including: RNA polymerase *β*- and *β′*-subunits [35] and group I sigma factor [18]. The other difference seen between the indel data and rRNA trees concerns the branching position of *T. maritima*. The rRNA phylogenies place this species in a distinct deep-branching group, whereas the indel data groups this species with other Gram-positive bacteria. Note that, although *T. maritima* (based on the absence of a large insert in Hsp70) has been grouped with the Gram-positive group, the signature sequences in ribosomal S12 protein and DNA gyrase A subunit indicate that it is distinct from both the traditional low G+C and the high G+C Gram-positive bacteria [19]. It is thus probable that *T. maritima* forms a separate, deep lineage within the Bacteria, showing a close affinity to the Gram-positive bacteria.

Phylogenetic analysis based on indel data complements the major limitations of the 16 S rRNA trees

An important point that emerges from these studies is that the evolutionary inferences based on indel data are not contradictory to those based on 16 S rRNA trees, but complement such studies in important respects. The two main recognized weaknesses of the rRNA phylogenies are: (1) it has proven difficult to define the main groups within the Bacteria in clear molecular terms and (2) the rRNA trees cannot resolve the relative

Fig. 12 Alignment of phosphoribosylpyrophosphate synthetase, showing a 1-amino-acid insert (*boxed*) distinctive of β- and γ-proteobacteria

```
                                             75                                  131
             E. coli K12                     DALRRASAGRITAVIPYFGYARQDRRVRS A RVPITAKVVADFLSSVGVDRVLIVDLH
             E. coli O157:H7                 ----------------------------- - ---------------------T----
             E. coli O157:H7EDL933           ----------------------------- - ---------------------T----
             Buc. sp. APS                    -S--------------------------- - ---------------I------T----
γ-Proteo     X. fastidiosa                   ---K---VSSV---------S-----M-- L --------A-KMI-AI-A----TI---
             Hae. influenzae                 ----------------------------- - -----------L--I--I----TC---
             Pse. aeruginosa                 --F--S--T-----------------P-- - --A-S------M-TV---N---T----
             Vib. cholerae                   --M-------------------------- - --------------N-------TI---
             Past. multocida                 ----------------------------- - ---------------------TC---
β-Proteo     Nei. meningitidis Z2491         ---K--------TA------------P-- V ----S--L--NM-Y-A-I----T----
             Nei.meningitidis MC58           ---K--------TA------------P-- V ----S--L--NM-Y-A-I----T----
α-Proteo     C. crescentus                   ---K---GK----------------KTGG   -T--S--L--NLITRS-A----TM---
             Meso. loti                      --FM-S--K-----------------ASG   -T--S--L--NMITRA------TL---
             Camp. jejuni                    -----S--NS---I------------KANP   -------L--NLIQAA-I---ATI---
ε-Proteo     Hel. pylori 26695               -----S--NS----L-----------KAAP   -------M--NLMQE--IE-IITM---
             Hel. pylori J99                 -----S--NS----L-----------KAAP   -------M--NLMQE--IE-IITM---
             Aqu. aeolicus                   --VK-S-PKE----V--YA-G----QDKP   -T--S--L---LIQKA-AN--IV----
Spirochetes  Trep. pallidum                  --V-H-G---V-L-L-TYP-S--HKK CG   -EGL--GLLGSVYEYL--SHIVTL---
             Bor. burgdorferi                --CMQ-K-NSVSVI--SYP-S---KKHS    -ECL--SLIGR--EEL-IRHI-TL-I-
             Sy. sp.  PCC6803                --C-----RQ----L--Y----A--KTAG   -ES-S--L--NLITGA-AQ---AM---
             D. radiodurans                  --AKS-----V------YS---S-KKDSP   -IS-AGRL---L-QEA-A----TMT--
             T. maritima                     --FK----NT-AV----Y-------KAKG   -D--S--L--NLITVA-AT---T----
             Myc. tuberculosis CDC1551       ---K-G--K-----M-FYP-----KKH-G   -E--S-RLI--L-KTA-A--IVT----
High G+C     Myc. leprae                     ---K-G--K-------FYP-----KKH-G   -E--S-RL---L-KTA-A--IVT----
             Myc. tuberculosis H37Rv         ---K-G--K-----M-FYP-----KKH-G   -E--S-RLI--L-KTA-A--IVT----
             Strep. pyogenes                 ---K----EK-SV-M--Y-------KA--   -E---S-L--NM-EVA----L-T----
             Bac. subtilis                   ---K----KT-NI----Y-------KA--   -E-----LF-NL-ETA-AT--IAL---
             Bac. halodurans                 --VK----KT-NV----Y-------KA-A   -E-----L--NL-ETA-AT---TL---
             Lac. lactis                     --------AS-NI-L--Y-------KA-A   -E---S-L--NM-QIA-A--LITF---
Low G+C      Sta. aureus                     --CK----AT-NI-V--Y-------KA--   -E-----L--NLIETA-AT-MIAL---
             Sta. aureus NCTC 8325           --CK----AT-NI-V--Y-------KA--   -E-----L--NLIETA-AT-MIAL---
             M. pulmonis                     -S------KT-NVILS-Y-------KAEG   -Q--A--LL--L-QVA-IS-IVV----
             U. urealyticum                  -SIK----KA-SV----Y-------KAKP   -E----RL--KMIE-A-ATS--TW-I-
             M. pneumoniae                   ---K-G--KS---IL--Y-------KTMG   -E---S-L---L-TTA--S--ALT-I-
             M. genitalium                   ---K-G--KS---IL--Y-------KTKG   -E---S-LI--M-TKA-AN--VLT-I-
```

Fig. 13 Sequence alignment of 5′-phosphoribosyl aminoimidazole-4-carboxamide transformylase, showing a 2-amino-acid deletion that is distinctive of the γ-proteobacteria

```
                                                  59                                   108
             E. coli K12                          GFPEMMDGRVKTLHPKVHGGILGRRGQDDAI    MEEHQIQPIDMVVVNLYPF
             E. coli O157                         -------------------------------    -------------------
             E. coli O157:H7EDL933                -------------------------------    -------------------
             Pse. aeruginosa                      -----------------------------G-    -AQ-G-----I--------
γ-Proteo     Vib. cholerae                        ------------------------------V    -NT-G--------------
             Pas. multicoda                       --------------------------T--EV    -SQQG-EG-----------
             Xylella fastidiosa                   ----------------M----L---A-I---V    -AK-G-A---LLIL-----
             Buchnera sp. APS                      ---I-----------IM-----QKQK-QE-    -KLYN-C---I-I--F---
             H. influenzae                        --------------------------T----    -QQ-G-EG-----------
β-Proteo     Nei. meningitidis MC58               -----L----------I--------DLPEHV AK ----G-GN--L-C------
             Nei. meningitidis Z2491              -----L----------I--------DLPEHV AK ----G-GN--L-C------
             C. crescentus                        ----------------V----L--V-DAA-HA KA -AD-G-GG--ILY------
α,ε-Proteo   Mesorhizobium loti                    ---I----------S---AL--V-DDPEHA AA -RKYG-E---LL-S-----
             Camp. jejuni                         ----LFE---------I-----HK-SDENH- KQ AK-NEXLG--L-C------
             Aqu. aeolicus                        ----ILE--------V------F-DWVEKDK EE I-K-G-K---V--------
             Sy. sp. PCC6803                      -A--ILG--------RI-----A--DLPSDQ AD L-AND-R-L-L--------
             D. radiodurans                       -----L---------AI-----A--EAG HL GQ LAAQD-GT--L-C------
             T. maritima                          --ENLLG-L------EIFA----PEPR              W-V-F-D---P
             Myc. leprae                          ----VL---------R--A-L-ADLRKPEHA AA L-QLG-EAFEL--------
High G+C     Myc. tuberculosis H37 Rv             ----VL---------R--A-L-ADLRKSEHA AA L-QLG-EAFEL--------
             Myc. tuberculosis CDC1551            ----VL---------R--A-L-ADLRKSEHA AA L-QLG-EAFEL--------
             Bac. subtilis                        ----I----L-----NI---L-AV--NEEHM AQ IN--G-----L--------
             Bac. halodurans                      ----IL---------NI---L-AM-ER-EHL AQ LN--H-R---F--------
Low G+C      Lactococcus lactis                   ---------L-----LI--AL----DLESHM KS -T--H-S---L--------
             Sta. aureus N315                      ---I-----------A------AD-NKPQHL NE LS-QH-DL-----------
             Sta. aureus MU50                      ---I-----------A------AD-NKPQHL NE LS-QH-DL-----------
             Strep. pyogenes                       ---------------NI---L-A--DA-SHL QA AKDNN-EL--L--------
```

branching order of the main groups. These are in fact strong points for the signature sequence approach. The main reason for the success of the signature sequence approach in these regards is that the derived inferences in this case are based on minimal assumptions [19, 56, 58, 63]. The sole assumption involved in these analyses is that when a shared conserved indel is present in different groups of species, it is assumed to have been introduced only once in a common ancestor of these groups, rather than on multiple occasions in different species. This is the most parsimonious way to explain these results. In contrast, the branching patterns of species in phylogenetic trees are dependent upon and affected by a large number of variables and assumptions (e.g., sequence regions that are retained or excluded, the number and range of species examined, differences in the evolutionary rates between species, base compositional differences between species, phylogenetic methods employed, order in which different species are added to the alignment, etc.) and hence are not clearly resolved [19, 38, 70].

Based on the various indels described here, it is now possible to define in clear molecular terms most of the major groups within the Bacteria that were previously

Table 2 The distribution of various indels in different proteins from bacterial genomes

Protein	Signature description	No. of genomes with protein	Genomes lacking the protein	No. of genomes with insert (expected/found)	No. of genomes lacking the insert (expected/found)	Exceptions observed
Hsp70/DnaK	21–23-a.a. G+/G– insert	41	None	27/27	14/14	0
Ribosomal S12 protein	13-a.a. low G+C signature	41	None	37/73	31/31	0
Hsp60/GroEL	1-a.a. insert after *Deinococcus*	39	mp, uu	26/26	37/68	0
FtsZ protein	1-a.a. insert after cyanobacteria	33	ct, cp, cm, mn, mg, uu	20/20	37/68	0
Ata-tRNA synthetase	4-a.a. common to *Chlamydia*/proteobacteria	41	None	23/23	18/18	0
Hsp70/DnaK	2-a.a. proteobacterial insert	41	None	17/17	24/24	0
CTP Synthetase	10-a.a. proteobacterial insert	37	mp, mg, uu, mn	17/17	20/20	0
Lon protease	1-a.a. $\alpha\beta\gamma$-proteobacterial deletion	33	ll, mt, ml, sa, sp	19/19	14/14	0
SecA protein	7-a.a. $\alpha\beta\gamma$-proteobacterial insert	41	None	14/14	27/27	0
HSP70/DnaK	4-a.a. $\beta\gamma$-proteobacterial insert	41	None	37/05	30/30	0
PRPP synthetase	1-a.a. $\beta\gamma$-proteobacterial insert	35	cp, ct, cm, rp	37/05	24/24	0
PAC-transfor mylase	2-a.a. γ-proteobacterial deletion	27	bb, cp, cm, ct, hp, mp, mg, tp, uu, rp	18/18	37/42	0

The abbreviations used are: a.a., amino acid; bb, *Borrelia burgdorferi*; cm, *Chlamydia muridarum*; cp, *Chlamydia pneumonia*; ct, *Chlamydia trachomatis*; G+, Gram-positive; G–, Gram-negative; hp, *Heliobacter pylori*; ll, *Lactococcus lactis*; mg, *Mycoplasma genitalium*; mn, *M. pneumonia*; mp, *M. pulmonis*; ml, *Mycobacterium leprae*; mt, *Myc. tuberculosis*; rp, *R. prowazekii*; sa, *Staphylococcus aureus*; sp, *S. pyogenes*; tp, *Treponema pallidum*; uu, *Ureaplasma urealyticum*

identified solely on the basis of their branching pattern in the 16 S rRNA trees. For example, the low G+C Gram-positive group can be defined by the presence of the large insert in the S12 protein. The high G+C Gram-positive group can be defined by the lack of the large inserts in both the Hsp70 protein and the S12 protein. A flow chart detailing how these indels could be used to taxonomically define the different main groups within the Bacteria and for assigning any given species to one of these groups has been described in earlier work [23]. The branch orders of different groups as deduced, based on these signatures, is internally highly consistent and it is difficult to explain these results by any other reasonable mechanism [63]. It should be recognized, however, that the number of main groups within the Bacteria that can presently be identified by signature sequence represents the minimal number. As additional signature sequences are identified in future work, the relative branching orders of species within some of the presently defined groups should become clearer; and this may lead to further divisions of these groups. We expect this to be the case for the low and high G+C Gram-positive bacteria and for the *Aquifex*, *Chlamydiae*, and *Cytophaga* groups, which have not been studied in detail for the presence of signature sequences. It is expected, however, that any newly identified group should be placed in an adjoining position to the presently assigned position and it should not affect the overall branching order of the other groups.

Conclusions

Results presented here show that the conserved indels that have been identified in various proteins provide a powerful new approach for understanding bacterial phylogeny. Based on these signatures, most of the main groups within the Bacteria can be identified in clear molecular terms and any given bacterial species could be assigned to one of these groups in an unambiguous manner. The phylogenetic assignment of different bacteria whose genomes have been sequenced using this approach showed an excellent correlation to that based on the 16 S rRNA, with 39 of the 41 species similarly assigned. Thus, the inferences deduced based on this new approach are not contradictory to the 16 S rRNA trees, but complement it in important respects. One distinct advantage of this new approach is that it permits a logical deduction of the relative branching order of different groups of bacteria from a common ancestor (Fig. 1), which could not be resolved from phylogenetic trees based on the 16 S rRNA or various proteins and constituted a major unresolved problem in bacterial phylogeny. The deduced branching order of different groups shows a very high degree of internal consistency and it is strongly supported by the analyses of completed bacterial genomes. As sequence information from other bacterial genomes becomes available, it should be possible to further determine: (1) whether the results obtained are in accordance with this model and (2) the ability of this model to help explain and integrate different observations.

Appendix

The following figures illustrate the alignment of various protein sequences from completed bacterial genomes, as discussed in the text.

The first five groups of proteins in this Appendix cover the ribosomal S12 protein (2), Hsp70 protein (3), Hsp60/GroEL protein (4), FtsZ protein (5), and alanyl-tRNA synthetase (6).

The remaining seven groups cover signature sequences for proteobacteria in Hsp70 and CTP synthase (7, 8), signature sequences indicating the branch order of the proteobacterial groups (9, 10), and useful signatures for the β- and γ-proteobacteria in the Hsp70 family of proteins (11, 12, 13).

References

1. Alm RA, Ling LSL, Moir DT, et al (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397:176–180
2. Andersson SGE, Zomorodipour A, Andersson JO, et al (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140
3. Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. Trends Genet 14:442–444
4. Balows A, Trüper HG, Dworkin M, Harder W, Schleifer KH (1992) The prokaryotes. Springer, Berlin Heidelberg New York
5. Blattner FR, Plunkett G III, Bloch CA, et al (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1462
6. Brown JR, Doolittle WF (1997) Archaea and the prokaryote-to-eukaryote transition. Microbiol Rev 61:456–502
7. Chambaud I, Heilig R, Ferris S, et al (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res 29:2145–2153
8. Cole ST, Brosch R, Parkhill J, et al (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–538
9. Cole ST, Eiglmeier K, Parkhill J, et al (2001) Massive gene decay in the leprosy bacillus. Nature 409:1007–1011
10. Deckert G, Warren PV, Gaasterland T, et al (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature 392:353–358
11. Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124–2128
12. Ferretti JJ, McShan WM, Ajdic D, et al (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci USA 98:4658–4663
13. Fleischmann RD, Adams MD, White O, et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512
14. Fraser CM, Casjens S, Huang WM, et al (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580–586
15. Fraser CM, Gocayne JD, White O, et al (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403
16. Fraser CM, Norris SJ, Weinstock CM, et al (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388
17. Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 407:757–762
18. Gruber TM, Bryant DA (1998) Characterization of the group 1 and group 2 sigma factors of the green sulfur bacterium *Chlorobium tepidum* and the green non-sulfur bacterium *Chloroflexus auranticus*. Arch Microbiol 170:285–296
19. Gupta RS (1998) Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes. Microbiol Mol Biol Rev 62:1435–1491
20. Gupta RS (1998) What are archaeobacteria: 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
21. Gupta RS (2000) Evolutionary relationships among *Bacteria*: does 16 S rRNA provide all the answers? ASM News 66: 189–190
22. Gupta RS (2000) The natural evolutionary relationships among prokaryotes. Crit Rev Microbiol 26:111–131
23. Gupta RS (2000) The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol Rev 24:367–402
24. Gupta RS, Golding GB (1993) Evolution of HSP70 gene and its implications regarding relationships between archaeobacteria, eubacteria, and eukaryotes. J Mol Evol 37:573–582
25. Gupta RS, Singh B (1992) Cloning of the HSP70 gene from *Halobacterium marismortui*: relatedness of archaeobacterial HSP70 to its eubacterial homologs and a model for the evolution of the HSP70 gene. J Bacteriol 174:4594–4605
26. Gupta RS, Mukhtar T, Singh B (1999) Evolutionary relationships among photosynthetic prokaryotes (*Heliobacterium chlorum*, *Chloroflexus aurantiacus*, cyanobacteria, *Chlorobium tepidum* and proteobacteria): implications regarding the origin of photosynthesis. Mol Microbiol 32:893–906
27. Heidelberg JF, Eisen JA, Nelson WC, et al (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature 406:477–483
28. Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 24:4420–4449
29. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359
30. Jain R, Rivera M, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806
31. Kalman S, Mitchell W, Marathe R, et al (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat Genet 21:385–389
32. Kaneko T, Nakamura Y, Sato S, et al (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res 7:331–338
33. Kaneko T, Sato S, Kotani H, et al (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
34. Kawula TH, Lelivelt MJ (1994) Mutations in a gene encoding a new Hsp70 suppress rapid DNA inversion and *bgl* activation, but not *proU* derepression, in *hns-1* mutant *Escherichia coli*. J Bacteriol 176:610–619
35. Klenk HP, Meier TD, Durovic P, et al (1999) RNA polymerase of *Aquifex pyrophilus*: Implications for the evolution of the bacterial *rpoBC* operon and extremely thermophilic bacteria. J Mol Evol 48:528–541

36. Kunst F, Ogasawara N, Moszer I, et al (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature 390:249–256
37. Kuroda M, Ohta T, Uchiyama I, et al (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 357:1225–1240
38. Lake JA (1991) The order of sequence alignment can bias the selection of tree topology. Mol Biol Evol 8:378–385
39. Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci USA 95:9413–9417
40. Ludwig W, Klenk H-P (2001) Overview: a phylogenetic backbone and taxonomic framework for prokaryotic systematics. In: Boone DR, Castenholz RW (eds) Bergey's manual of systematic bacteriology. Springer, Berlin Heidelberg New York, pp 49–65
41. Ludwig W, Schleifer KH (1999) Phylogeny of *Bacteria* beyond the 16 S rRNA standard. ASM News 65:752–757
42. Maidak BL, Cole JR, Lilburn TG, et al (2001) The RDP-II (ribosomal database project). Nucleic Acids Res 29:173–174
43. Makino K, Yokoyama K, Kubota Y, et al (1999) Complete nucleotide sequence of the prophage VT2-Sakai carrying the verotoxin 2 genes of the enterohemorrhagic *Escherichia coli* O157:H7 derived from the Sakai outbreak. Genes Genet Syst 74:227–239
44. May BJ, Zhang Q, Li LL, Paustian ML, Whittam TS, Kapur V (2001) Complete genomic sequence of *Pasteurella multocida*, Pm70. Proc Natl Acad Sci USA 98:3460–3465
45. Morowitz HJ (1992) Beginnings of cellular life: metabolism recapitulates biogenesis. Yale University Press, New Haven
46. Murray RGE (1986) Family II. *Deinococcaceae* Brooks and Murray 1981, 356^{VP} In: Sneath PHA, Mair NS, Sharpe ME, Holt JG (eds) Bergey's manual of systematic bacteriology. Williams and Wilkins, Baltimore, pp 1035–1043
47. Murray RGE (1986) Kingdom procaryote. In: Sneath PHA, Mair NS, Sharpe ME, Holt JG (eds) Bergey's manual of systematic bacteriology. Williams and Wilkins, Baltimore, pp 34–36
48. Murray RGE (2000) Evolutionary relationships and taxonomy. ASM News 66:324–325
49. Nelson KE, Clayton R, Gill S, et al (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. Nature 399: 323–329
50. Nierman WC, Feldblyum TV, Laub MT, et al (2001) Complete genome sequence of *Caulobacter crescentus*. Proc Natl Acad Sci USA 98:4136–4141
51. Olsen GJ, Woese CR, Overbeek R (1994) The winds of (evolutionary) change: breathing new life into microbiology. J Bacteriol 176:1–6
52. Parkhill J, Achtman M, James KD, et al (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature 404:502–506
53. Parkhill J, Wren BW, Mungall K, et al (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403:665–668
54. Perna NT, Plunkett G III, Burland V, et al (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529–533
55. Read TD, Brunham RC, Shen C, et al (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 28:1397–1406
56. Rivera MC, Lake JA (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. Science 257:74–76
57. Seaton BL, Vickery LE (1994) A gene encoding a DnaK/hsp70 homolog in *Escherichia coli*. Proc Natl Acad Sci USA 91: 2066–2070
58. Sekowska A, Danchin A, Risler JL (2000) Phylogeny of related functions: the case of polyamine biosynthetic enzymes. Microbiology 146:1815–1828
59. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature 407:81–86
60. Shirai M, Hirakawa H, Kimoto M, et al (2000) Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. Nucleic Acids Res 28: 2311–2314
61. Simpson AJG, Reinach FC, Arruda P, et al (2000) The genome sequence of the plant pathogen *Xylella fastidiosa*. Nature 406:151–157
62. Stephens RS, Kalman S, Lammel C, et al (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science 282:754–759
63. Stiller JW, Hall BD (1999) Technical comments: http://www.sciencemag.org/cgi/content//full/286/5444/1443a. Science 286:1443a
64. Stover CK, Pham XQ, Erwin AL, et al (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. Nature 406:959–964
65. Takami H, Nakasone K, Takaki Y, et al (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res 28:4317–4331
66. Tettelin H, Saunders NJ, Heidelberg J, et al (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science 287:1809–1815
67. Tomb JF, White O, Kerlavage AR, et al (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388:539–547
68. White O, Eisen JA, Heidelberg JF, et al (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286:1571–1577
69. Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271
70. Woese CR (1991) The use of ribosomal RNA in reconstructing evolutionary relationships among bacteria. In: Selander RK, Clark AG, Whittmay TS (eds) Evolution at molecular level. Sinauer, Sunderland, pp 1–24
71. Woese CR (1998) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859
72. Woese CR, Stackebrandt E, Macke RJ, Fox GE (1985) A phylogenetic definition of the major eubacterial taxa. Syst Appl Microbiol 6:143–151