Phylogenetic, amino acid content and indel analyses of the β subunit of DNA-dependent RNA polymerase of Gram-positive and Gram-negative bacteria

Food Microbial Sciences
Unit, School of Food
Biosciences, University of
Reading, Whiteknights,
PO Box 226, Reading
RG6 6AP, UK

Robert Morse, Karen O'Hanlon and Matthew D. Collins

Author for correspondence: Matthew D. Collins. Tel: +44 118 9357226. Fax: +44 118 9357222.
e-mail: m.d.collins@reading.ac.uk

In this study, we have sequenced the *rpoB* gene, encoding the β subunit of DNA-dependent RNA polymerase, from a selection of Gram-positive and Gram-negative bacteria. The presence of insertions and deletions (indels) in the β subunit separate the Gram-positive and Gram-negative bacteria from each other and support the division of the Gram-positive organisms into two clades based on DNA G+C content. Phylogenetic and amino acid content analyses further separate the clostridia from bacilli, leuconostocs, listeriae and relatives, forming an early branch after the common Gram-positive ancestor. The occurrence in the β subunit of Asn-Ala at positions 471–472 in *Porphyromonas gingivalis* and Asn at position 372 in *Weissella paramesenteroides* are postulated to be the cause of the natural rifampicin resistance of these species.

Keywords: phylogeny, Gram-positive bacteria, DNA-dependent RNA polymerase, *rpoB* gene, rifampicin resistance

INTRODUCTION

Due to the pioneering work of Carl Woese (Woese, 2000), 16S rRNA has increasingly occupied the centre stage in microbial systematics and has revolutionized our understanding of prokaryote evolution and diversity. Until recently, nearly all phylogenetic reconstructions amongst the bacteria were based on this single molecular chronometer. Recently, however, systematists have begun to compare and contrast phylogenetic relationships inferred from other molecules with those of 16S rRNA (Gupta, 1998; Hansmann & Martin, 2000). Examples of such alternative chronometers include GrpE (Ahmad *et al.*, 2000), ribosomal protein (Bocchetta *et al.*, 2000), Hsp70 (Gribaldo *et al.*, 1999), Hsp40 (Bustard & Gupta, 1997) and EF-G (Bocchetta *et al.*, 1995). These studies have, with a few exceptions (Hasegawa & Hashimoto, 1993), demonstrated that 16S rRNA is an excellent

molecule for reconstructing phylogenetic relationships, although recent evidence of lateral transfer of rRNA gene segments suggests that caution should always be observed in interpreting such analyses (Wang & Zhang, 2000).

DNA-dependent RNA polymerase (RNAP) is a multi-subunit enzyme consisting of two α subunits (encoded by the *rpoA* gene), one β subunit (*rpoB*) and one β' subunit (*rpoC*). The large sizes of the β and β' subunits (respectively 151 and 156 kDa in *Escherichia coli*), together with their ancient origin in evolutionary terms, indicate that these molecules could be immensely powerful molecular chronometers for phylogenetic analysis (Rowland *et al.*, 1992; Klenk & Zillig, 1994; Morse *et al.*, 1996a; Mollet *et al.*, 1997, 1998; Klenk *et al.*, 1999; Bocchetta *et al.*, 2000). The key role for RNAP in transcription of DNA makes this independent of the translational apparatus. Although natural intraspecific horizontal *rpoB* gene fragment exchange has been demonstrated (Lorenz & Sikorski, 2000), this does not necessarily preclude its use for phylogenetic analysis.

One of the earliest weighted characters employed in bacterial classification was the Gram stain, which, although now thought unsuitable for phylogenetic

A multiple alignment of the RNAP β subunit amino acid sequences deduced from the gene sequences is available as supplementary material in IJSEM Online (<http://ijs.sgmjournals.org/>).

Abbreviation: RNAP, DNA-dependent RNA polymerase.

The EMBL accession numbers for the *rpoB* gene sequences reported in this study are Y16466–Y16472, as listed in Table 1.

Table 1. Strains used and their *rpoB* gene sequence accession numbers

Strain	<i>rpoB</i> gene accession no.
<i>Clostridium argentinense</i> ATCC 27322 ^T (toxin type G)	Y16472
<i>Leuconostoc mesenteroides</i> NCIMB 8023 ^T	Y16467
<i>Clostridium botulinum</i> NCTC 7272	Y16466
<i>Listeria monocytogenes</i> NCTC 7973	Y16468
<i>Listeria murrayi</i> NCTC 10812 ^T	Y16469
<i>Weissella paramesenteroides</i> NCIMB 13092 ^T	Y16471
<i>Porphyromonas cangingivalis</i> NCTC 4874 ^T	Y16470

analysis, has become an essential dichotomy for classification schemes (Shah *et al.*, 1997). The application of 16S rRNA analyses to relationships within the Gram-positive bacteria has produced differing patterns of relationships that depend on the method of analysis used (Olsen *et al.*, 1994; Ludwig *et al.*, 1994) and which are sometimes in conflict with those defined either by other gene products, e.g. EF-Tu (Ludwig *et al.*, 1994), ATPase β (Ludwig *et al.*, 1994) and RecA (Eisen, 1995), or by whole genome analysis (Tekaia *et al.*, 1999). The classification of Gram-positive bacteria into two groups based on DNA G+C content (high and low G+C) is, however, supported by all these chronometers, together with studies on the types of RNaseP (Haas *et al.*, 1996) and DNA polymerase C (Huang & Ito, 1999) that they contain and the organization of their *dnaK* operons (Segal & Ron, 1996). Inferences within these groups are again varied, with phylogenetic studies on the HrcA (Ahmad *et al.*, 1999) and GrpE (Ahmad *et al.*, 2000) proteins suggesting two separate groups of low-G+C Gram-positive bacteria. The relationship of these two (three) Gram-positive groups with the Gram-negative bacteria is also unclear, with the high-G+C group being more closely linked on the basis of RNaseP RNA type (Haas *et al.*, 1996) and the presence of shared indels in S12 protein, dihydroorotate dehydrogenase and pyruvate kinase (Gupta, 1998), whilst the low-G+C group is more closely linked on the basis of shared indels in gyrase A (Gupta, 1998). Phylogenetic trees based on various chronometers have indicated close relationships of Gram-positive bacteria to various groups of other bacteria including the cyanobacteria (Ahmad *et al.*, 1999; Eisen, 1995), halobacteria (Gribaldo *et al.*, 1999; Haas *et al.*, 1996) and mycoplasmas (Weisburg *et al.*, 1989a), whilst specific links have been suggested to *Methanosarcina mazei* (Gribaldo *et al.*, 1999) and *Borrelia burgdorferi* (Eisen, 1995). Likewise, such chronometers also give variable phylogenetic positions for the clostridia (Ahmad *et al.*, 1999, 2000; Olsen *et al.*, 1994; Gribaldo *et al.*, 1999; Bustard & Gupta, 1997).

Previously, we have sequenced the *rpoC* gene from various leuconostocs and performed a comparative phylogenetic analysis to discern relationships and the tenet of tachytelec evolution within this group of Gram-positive bacteria (Morse *et al.*, 1996a). In this

study, we have extended this work by sequencing the *rpoB* gene from various Gram-positive bacteria to determine the inter- and intrarelations of this group of bacteria.

METHODS

Bacterial strains and DNA preparation. The strains used in this study are listed in Table 1. Strains were cultured in appropriate media and at temperatures recommended in the relevant culture collection catalogues. Chromosomal DNA was extracted as described previously (Harland *et al.*, 1993). The authenticity of each strain examined was confirmed by partial 16S rRNA gene sequence analysis (data not shown).

PCR amplification. In order to obtain nearly complete copies of the *rpoB* gene, PCR products were generated using oligonucleotide primers (Table 2) complementary to DNA sequences encoding conserved amino acid sequences in the *rpoB* genes of *Mycobacterium tuberculosis* (EMBL accession no. U12205), *Mycobacterium leprae* (Z14314), *Mycobacterium smegmatis* (U24494), *Bacillus subtilis* (L24376) and *Staphylococcus aureus* (X64172) and the *rplL* genes of *Escherichia coli* (V00339) and *Staphylococcus aureus* (X64172). PCRs were performed in 1 × PCR buffer (Perkin-Elmer) containing 10 mM dNTPs and 4 ng of each oligonucleotide primer μl^{-1} in a final volume of 50 μl under a layer of PCR-grade mineral oil (Sigma) using a Biometra PCR thermal cycler. After an initial denaturation step of 94 °C for 5 min, AmpliTaq polymerase (Perkin-Elmer) was added to a final concentration of 0.02 U μl^{-1} and 25 cycles of 92 °C for 1 min, 48 °C for 1 min and 65 °C for 3 min were performed, ending with a final extension step of 65 °C for 10 min.

Cloning, transformation and sequencing. PCR products were purified from primers, nucleotides and enzyme using a QIAquick PCR purification kit (Qiagen), ligated into pCRII vector (TA cloning kit; Invitrogen) and transformed into *Escherichia coli* INV α I cells according to the manufacturer's instructions. DNA sequencing was performed by using the dideoxynucleotide chain termination method on both positive and negative strands of each cloned PCR product. Two independent PCR products were sequenced for each *rpoB* gene to check AmpliTaq polymerase fidelity, with a third PCR product being generated to resolve any nucleotide base discrepancies. On average, one nucleotide base error was detected for every 3 kb of sequence.

Sequence analysis. Sequences were aligned using the PILEUP program from the Wisconsin Molecular Biology software package (Devereux *et al.*, 1984) and the multiple alignments were edited manually. Distance matrices were calculated for the DNA alignments using the programs PRETTY and

Table 2. Oligonucleotides used to generate PCR products

All sequences and positions are given 5' → 3'. Positions refer to the *Bacillus subtilis* *rpoB* gene or the *Staphylococcus aureus* *rplL* gene.

Primer	Sequence	Position
1475	GA(A/G)AA(A/G)ACNGA(A/G)TT(T/C)GA(T/C)GT	151–166
2233	GTNTT(T/C)ATGGGNGA(T/C)TT(T/C)CC	331–350
2234	GG(G/A)AA(G/A)TCNCCCAT(G/A)AANAC	350–331
2235	AA(G/A)(A/C)GNTA(T/C)GA(C/T)(C/T)TNGC	787–803
2236	GCNA(G/A)(G/A)TC(A/G)TANC(T/G)(C/T)TT	803–787
2237	GA(C/T)GA(C/T)ATNGA(C/T)CA(C/T)(C/T)T	1195–1211
2238	A(G/A)(G/A)TG(G/A)TCNAT(G/A)TC(G/A)TC	1211–1195
2239	GA(G/A)GTN(C/A)GNGA(C/T)GTNCA	1504–1520
2240	TGNAC(G/A)TCNC(G/T)NAC(C/T)TC	1520–1504
2074	AA(C/T)ATGCAA(G/A)(C/A)GNCA(G/A)GCNGT	1924–1943
2243	ACNGC(C/T)TGNC(G/T)(C/T)TGCAT(G/A)TT	1943–1924
2241	GA(G/A)GA(G/A)ATNACNCGNGA(T/C)AT	2419–2438
2242	AT(A/G)TCNCGNGTNAT(C/T)TC(C/T)TC	2438–2419
2076	GCNATNTT(T/C)GGNGA(G/A)AA(G/A)GC	2587–2606
2244	GC(C/T)TT(C/T)TCNCC(A/G)AANATN(G/T)C	2606–2587

DNADIST (using the Kimura two-correction parameter) (Felsenstein, 1989) and for the amino acid alignment using the GeneBase program (Applied Maths). Phylogenetic trees were generated using the neighbour-joining method with the program NEIGHBOR and were then tested statistically by bootstrap analysis (Felsenstein, 1989). Bootstrap values were calculated for the DNA trees from 1000 replicates using the programs SEQBOOT, DNADIST, NEIGHBOR and CONSENSE (Felsenstein, 1989) and for the protein tree from 100 replicates using the GeneBase program.

RESULTS AND DISCUSSION

The PCR amplification strategy generated complete copies of the *rpoB* gene from the Gram-positive bacterium *Listeria monocytogenes* and the Gram-negative bacterium *Porphyromonas cangingivalis* and partial copies of the gene from *Clostridium argentinense* (encoding a region equivalent to positions 117–1068 of the *Bacillus subtilis* β subunit), *Clostridium botulinum* (113–1059), *Leuconostoc mesenteroides* (111–1081), *Weissella paramesenteroides* (117–1047) and *Listeria murrayi* (121 to the end of the subunit). The fidelity of the sequence data generated from PCRs was checked rigorously and discrepancies were resolved by sequencing additional products from independent PCRs. A multiple alignment of the β subunit amino acid sequences deduced from the gene sequences together with the equivalent regions from a selection of other bacteria is available as supplementary material in IJSEM Online (<http://ijs.sgmjournals.org/>). As is evident from the alignment, and as we have shown previously for the β' subunit (Morse *et al.*, 1996b), the numerous regions of high primary structural conservation that are interspersed throughout the β subunit molecule greatly facilitate the alignment of less-conserved regions. Exclusion of such regions due

to poor alignment can greatly affect inferred phylogenies (Hansmann & Martin, 2000). The natural rifampicin resistance of *Porphyromonas gingivalis* and *Porphyromonas cangingivalis* (Klimpel & Clark, 1990; Collins *et al.*, 1994) is possibly due to the sequence Asn–Ala found at positions 470–471 in the former (NCBI BLAST search of unpublished sequence) and 471–472 in the latter species (present work). Mutations at the equivalent highly conserved site Ser–Gln in other β subunits result in rifampicin resistance, such as Ser464Pro and Gln465Arg in *Staphylococcus aureus* (Aubry-Damon *et al.*, 1998; Wichelhaus *et al.*, 1999) and Ser–Gln (509–510) mutated to Val in *Escherichia coli* (Jin *et al.*, 1988). The natural rifampicin resistance of *Weissella paramesenteroides* (Swenson *et al.*, 1990) is possibly due to the presence of Asn at position 372, which is usually occupied by a highly conserved His in most other β subunits (Morse *et al.*, 1999). Mutation of this His to Asn induces rifampicin resistance in *Helicobacter pylori* (Heep *et al.*, 1999), *Streptococcus pneumoniae* (Enright *et al.*, 1998), *Staphylococcus aureus* (Wichelhaus *et al.*, 1999) and *Mycobacterium tuberculosis* (Telenti *et al.*, 1993). Asn also occurs at the equivalent position in the β subunit of the naturally rifampicin-resistant mycoplasmas (Liberal & Boughton, 1991) and *Spiroplasma citri* (Gaurivaud *et al.*, 1996). Another region of RNAP that has recently been identified as important to rifampicin binding is the F segment of the β' subunit (Naryshkina *et al.*, 2001). In particular, a stretch of four amino acids in this region (Thr–Ala–Leu–Lys at positions 786–789 in *Escherichia coli*) forms van der Waals interactions with amino acids in the β subunit (567–571 in *Escherichia coli*) to form the rifampicin-binding site. However, we have shown previously that these four amino acids are also found at the equivalent positions in the β' subunits of

Fig. 1. Comparison of unrooted neighbour-joining trees inferred from (a) 16S rRNA and (b) the amino acid sequence of the β subunit of RNAP. Bars, 2% sequence divergence.

Porphyromonas cangingivalis (EMBL accession no. Y16470) and *Weissella paramesenteroides* (Morse *et al.*, 1996a) and, therefore, cannot be the cause of the natural rifampicin resistance in these bacteria.

Obtaining good sequence alignments is a prerequisite of molecular systematic analysis because each aligned position has to include only homologous residues from the different molecules. However, several stretches of amino acid sequence were found only in the mycobacteria and Gram-negative species. In order to minimize alignment ambiguities due to these regions and to alleviate possible error in phylogenetic reconstruction due to incorrect weighting of such differences between species (e.g. residues 941–1040 in *Escherichia coli* and residues 283–361 in *Mycoplasma pneumoniae*), these indels were omitted. The reliability of phylogenetic inferences made from comparative analyses of 16S rRNA for a range of bacterial species was assessed by constructing neighbour-joining trees from the β subunit data. Trees were generated using both amino acid similarities and the equivalent DNA sequence at the second codon position. Studies using the B subunit (the equivalent of the bacterial β subunit) of archaeal RNAP have shown the nucleotide at the second codon positions to be unbiased by G+C contents and more reliable than positions 1 and 3 for phylogenetic reconstructions (Klenk & Zillig, 1994). Since differences in species sampling influence tree generation (Leconte *et al.*, 1993), a 16S rRNA tree was constructed using the same set of species to assess congruence between the two chronometers.

A comparison of the neighbour-joining trees derived from the β subunit and 16S rRNA analyses is shown in Fig. 1. The two β subunit trees, based on amino acid similarities and second codon position of *rpoB*, possess very similar topologies and the latter is therefore not

shown. All statistically significant groups (bootstrap values $\geq 90\%$) were confirmed by maximum-parsimony analysis (data not shown). The overall topology of the β subunit tree (Fig. 1b) is in good agreement with that constructed from 16S rRNA (Fig. 1a) with the exception of the positions of *Synechocystis* and the mycoplasmas. Discrepancies in the phylogenetic position of these organisms have been noted using other chronometers and ascribed to either variation in data selection or long-branch effects (Klenk *et al.*, 1999; Tekaiia *et al.*, 1999; Bocchetta *et al.*, 2000; Hansmann & Martin, 2000). Other minor topological differences between the trees invariably involve reference species (e.g. *Brochothrix thermosphacta*) where bootstrap values are low (data not shown) and are therefore of uncertain branching order. Little significance can therefore be attached to these differences. The β subunit tree shows the separation of the Gram-positive bacteria into two distinct clades representing the high-G+C bacteria (*Actinobacteria*) and the low-G+C bacteria. The latter group appears to be further subdivided into a clade containing clostridial species (26–28 mol% G+C) and one containing the listeriae, *Bacillus subtilis*, *Staphylococcus aureus* and the leuconostocs (37–42 mol% G+C), although additional representatives of these groups are needed to be confident of this division. The aforementioned divisions based on G+C content are reflected in the amino acid content of the β subunit proteins. It has been shown (Singer & Hickey, 2000), based on the amino acids encoded by GC- and AT-rich codons, that GC-rich coding DNA tends to produce proteins rich in the amino acids Gly, Ala, Arg and Pro, whilst AT-rich coding DNA favours the inclusion of the amino acids Phe, Tyr, Met, Ile, Asn and Lys. The ratio of the G+C content of the GC-rich-DNA-encoded amino acids to the AT-rich-DNA-encoded

(a)		(b)
HIGH G+C GRAM POSITIVE	E.c. IRNG-----KGEVDD 444	RKV--TDGVVT 600
	H.i. IRNG-----RGEVDD	RKV--VDGQVT
	N.m. LRNG-----HGEVDD	RRV--IDGKVT
	C.d. LHAGETMTSPTGEV-IPVETDD	RKV--ENGVLV
	M.t. LHEGQTMTVPGGVE-VPVETDD	RKV--VDGVVS
	M.l. LHEGQSTMTVPGGVE-VPVETDD	RKV--VDGVVS
	M.s. LHEGQSTMTVPGGVE-FAVESHD	RKV--VEGVVT
	A.m. LHAGEDKMTAANNTE-IPVETDD	RKV--VEGRVT
	T.w. LHLGFSEVAVLNSNKSVPISTDD	RRV--VNGKVT
	St.c. LHAGETETVGESGRE-IVVETDD	RKV--VEGQVT
LOW G+C GRAM POSITIVE	T.f. LHAGEKELVRPHGT--FPIEVDD	RKV--VDGRIT
	C.a. LTYG-----IGYIDD	RVVDKKTGKVT
	C.b. LSYD-----IGYKDD	RRIDPKNKRT
	L.me. LQEG-----VGHVDD	RRVDWTTTHKVT
	W.p. LPLG-----VGTDD	RRVDWTTTHKVT
	L.mo. LLHG-----VGDTDD	RRVDPETNRVT
	L.mu. LLHG-----VGDTDD	RRVDPETNRVT
	B.s. LLHG-----VGDTDD	RRVDPETGKVT
	S.a. LLSG-----IGYTDD	RKVDLDTTHAIT
(c)		
HIGH G+C GRAM POSITIVE	E.c. HLGMAAKGIGDKINAMLKQQQEVAKLREFIQRAYDLGADVRRQKVDLSTFSDEEVMRLAENLRKGMPIATPV 1186	
	H.i. HLGAAKIGIGDKINAMLKQQQEVAKLREFIQRAYDLGADVRRQKVDLSTFSDEEVMRLAENLRKGMPIATPV	
	N.m. HLGWAAKIGIGERIDRLKEQRKAGELREFLNR----LSRSGKKEDLSLTDEEIIELASNLRRKGASFASPV	
	C.d. HLGWLA--GWKIDTEDPANAELLKTLPELDYDVP-----AGSLTATPV	
	M.t. HLGWCAHS--GWKVDAAKGVDPWAARL--PDELLLEAH-----ANAIIVSTPV	
	M.l. HLGWVAKS--GWKIDVAGGIPDWAARL--PEELLHAA-----PNQIVSTPV	
	M.s. HLGWAKA--GWNIDVLAGVDPWAKSL--PEELYSAP-----ADSTVATPV	
	A.m. HLGWLASQ--GWTIE--GDPDWAKNL--SAELRDVA-----PGTNTATPV	
	T.w. HLGWISKQ--GWKIE--GDPDWAKD--IRVREAQ-----PDSRVSSPV	
	St.c. HLGWLASR--GWDVS--GLAEWAQRLQVIGADKVE-----PGTNVATPV	
LOW G+C GRAM POSITIVE	T.f. HLGWLAH--GWKVE--GDDAEWKRRRLRDIGAHEAP-----PNSKVATPV	
	C.a. HLGWAASE-----LGWHIATPV	
	C.b. HLGAAASK-----LGWHIATPV	
	L.me. HLGFAAKK-----LGIHVASPV	
	W.p. HLGMAARE-----LGIHVASPV	
	L.mo. HLGMAARA-----LGIHVATPV	
	L.mu. HLGMAARA-----LGIHVATPV	
	B.s. HMGMAARY-----LGIHTASPV	
	S.a. HLGMAAKN-----LGIHVASPV	

Fig. 2. Amino acid sequences encoded by selected regions of *rpoB* gene sequences demonstrating (a) an insert specific to the high-G + C Gram-positive bacteria, (b) an insert specific to the low-G + C Gram-positive bacteria and (c) a partial and a complete deletion in the high- and low-G + C Gram-positive bacteria, respectively. Abbreviations: E.c., *Escherichia coli*; H.i., *Haemophilus influenzae*; N.m., *Neisseria meningitidis*; C.d., *Corynebacterium diphtheriae*; M.t., *Mycobacterium tuberculosis*; M.l., *Mycobacterium leprae*; M.s., *Mycobacterium smegmatis*; A.m., *Amycolatopsis mediterranei*; T.w., *Tropheryma whipplei*; St.c., *Streptomyces coelicolor*; T.f., *Thermobifida fusca*; C.a., *Clostridium argentinense*; C.b., *Clostridium botulinum*; L.me., *Leuconostoc mesenteroides*; W.p., *Weissella paramesenteroides*; L.mo., *Listeria monocytogenes*; L.mu., *Listeria murrayi*; B.s., *Bacillus subtilis*; S.a., *Staphylococcus aureus*.

amino acids for the β subunits is 1:33 for *Mycobacterium tuberculosis*, 1:04–1:13 for listeriae and leuconostocs and 0:86–0:89 for the clostridial species examined.

Although the bifurcation separating the high- and low-G + C Gram-positive clades was not supported by high bootstrap values, this division was, however, demonstrated unequivocally by the presence of a 12–13 amino acid insert found only in the mycobacteria, *Tropheryma whipplei* (EMBL accession no. AF243072), *Amycolatopsis mediterranei* (AF242549), *Corynebacterium diphtheriae* (BLAST search), *Streptomyces coelicolor* (BLAST search) and *Thermobifida fusca* (BLAST search) (Fig. 2a) and a two amino acid insert found only in the low-G + C bacteria β subunits (Fig. 2b). On this basis, the Gram-positive-specific indels represent insertions that occurred after the divergence of the high- and low-G + C clades. Such a division has been demonstrated previously based on signature sequences only found in pyruvate kinase from low-G + C bacteria and gyrase A from high-G + C bacteria (Gupta, 1998). In each case, great care was taken that sufficient

flanking sequence homology was present to ensure positional and size equivalence of the inserts (Philippe *et al.*, 1999). The β subunits of Gram-negative bacteria are distinguished from those of Gram-positive bacteria by two rather less well-defined indels (Figs 2c and 3a). One of these (Fig. 2c) represents a sequence that has been partially deleted in mycobacteria, *Tropheryma whipplei* and *Amycolatopsis mediterranei* (*Escherichia coli* positions 1150–1177) and completely deleted in the low-G + C Gram-positive bacteria (*Escherichia coli* positions 1124–1177). The conserved sequence flanking the right-hand side of the deletion suggests that the partial deletion occurred first, before the separation of the high- and low-G + C bacteria, and that the second occurred only in the low-G + C bacteria. This supports an affinity between the mycobacteria and the Gram-negative bacteria indicated previously by work on the classification of RNaseP RNA (Haas *et al.*, 1996) and the presence of indels in S12 protein, dihydroorolate dehydrogenase and pyruvate kinase (Gupta, 1998). The other indel specific to the Gram-negative bacteria (Fig. 3a), including that found in the sequence of the β subunit of *Porphyromonas cangingivalis* reported here,

(a)		
C.a.	IVDVKVFTKEN-----	795
C.b.	IVDVKIPTREN-----	798
L.me.	VQNVRIYTPEN-----	784
W.p.	VQDVRIPTREN-----	799
L.mo.	VLDVKIFTREA-----	898
L.mu.	VLDVKIFTREA-----	775
T.a.	VVGRRLRRGDP-----	811
T.m.	VIRVDVYDQND-----	992
E.c.	VIDVQVFTRDGVEK-DKRAL-EIEEMQ-----	952
H.i.	VIDVQVFTRDGVEK-DKRAL-EIEEMQ-----	953
N.m.	VIDVQVFTREGIQR-DKRAQS-IIDSE-----	976
P.c.	VIKTHLFSKAMHSKKDKGGVREIVKKL-----	922
A.p.	VIDVQVFTRRGG-KAAKDLAQKVERE-----	1037
A.a.	VIDVQVFTRRGG-TGKKDMLAEKVERE-----	1036
R.a.	DLNKHYYTTSNLQKSTESLLSYIYKEEKIRLMFLSYLFFIFKNKYKYLGIKKKIILDDLNIHLNQKMNWSFSSFYTKTILYR	1027
R.p.	VVEVRIFSRRGVEK-DQRAIA-IEKQQ-----	969
E.c.	-----LQAKKDLSEELQILEAGLFSRIRAVLVAGGVEAEKLDKLP-----	993
H.i.	-----LREAKKDLTEELEILEAGLFARVRNLLISSGADAAQLDKLD-----	994
N.m.	-----LKKYRLDLNDQLRIFDNDADFRIERMIVGQKANGGPMK-LAKGSEITTEYLAGLPS-----	1031
P.c.	-----DEEMEEKLAELEMLKKLIQLTDNKLNGIRNFEDTEIVPGVKLTTPAVCSKIEFGEV-----	981
A.p.	-----ELEALEREL-EKKKNLLMSGVEKVLKGLILGRKTKDKDKVIGKKEIKKGTEITEEI-FKEY-----	1095
A.a.	-----ELEALEQEL-EKKKNLLIITGRDKVLKGLVILGRKVEKDAKVGKTKVKKGTVIDEKV-FEEF-----	1094
R.a.	NNLNKILNFDLIYKMYLLICYKLSIEIFLKNKKDLKKECI-KQDNFKYVDIRFKNMYREIYKRNKKNDVLLNLIKMSLSK	1110
R.p.	-----IEKLAKDRDDELEIIEHFVFSWLEKLLVGHVIIN-----GPKQITAGQTITTEMLKGLSK	1024
C.a.	-----ADELPPGVNQLVRC	809
C.b.	-----GDELPPGVNKLVR	812
L.me.	-----GDELAPGVNMVVRV	798
W.p.	-----GDELSPGVNMVVRV	793
L.mo.	-----GDELPPGVNQLVRV	912
L.mu.	-----GDELPPGVNQLVRV	789
T.a.	-----GVELKPGVREVRV	825
T.m.	-----IAELGAGVLKVRV	1006
E.c.	---RDRWLEGLTDEEKQNLQELAEQYDELKHEFEKKLEAKRRKITQ-----	1052
H.i.	---RTKWLEQTIADDEKQNLQELAEQYELRKEFEHKLKVRKKIIQ-----	1053
N.m.	---RHDWFDIRLTDEDLAKQLELKVSLQKREAEDELYEIKKKLTK-----	1090
P.c.	---AIFNWTGDEHTDE-LVAKLISNYLRKA-KEIESE--YRRKKFDATI-----	1037
A.p.	SQYILARPENFFTDEELIKKIREIYTRSRQIDMLQKVYQEKKETLLK-----	1157
A.a.	VNYILGRPENFFEDDLIRKIREIAERTRTQIEMLNKVYEEKKETLLK-----	1156
R.a.	VFFIDLKKKIPYSMEFL-LKNFILQEMYILLFLYYIKEYNQNFVNFENDQIFVQKTKSINKKKDCLKNNSLRMVKV	1188
R.p.	GQLW---QITVEDANVMNEIEQIKIHYDEKKYALDKRFTTKVEKLQS-----	1082
(b)		
R.a.	KIRLMFLSYLFFIFKNKY-KYDLGIKKKIILDDLNIH	1008
R.a.	LNQKMNWSFSSFY--TKTILYRNLNKILNFDLIYKMYLLICYK	1052
R.a.	LNKMSLSKVFVFDLKKKIPYSMEFLK-NFILQE-MYILLFLYY	1144

Fig. 3. Amino acid sequences encoded by selected regions of *rpoB* gene sequences demonstrating (a) an insert specific to the Gram-negative bacteria and (b) an alignment showing how a region of the latter insert has been triplicated in *Reclinomonas americana*. Abbreviations (in addition to those listed in the legend to Fig. 2): T.a., *Thermus aquaticus*; T.m., *Thermotoga maritima*; P.c., *Porphyromonas cangingivalis*; A.p., *Aquifex pyrophilus*; A.a., *Aquifex aeolicus*; R.a., *Reclinomonas americana*; R.p., *Rickettsia prowazekii*.

has no homologue in the subunit from *Thermotoga maritima*. This is consistent with the inference from indels found in Hsp70 and glutamine synthase that this bacterium exhibits a resemblance to Gram-positive bacteria, although this is discordant with work based on signature sequences in pyruvate kinase and gyrase A, which indicates that it is Gram-negative (Gupta, 1998). However, a homologous sequence is found in the equivalent subunit of the bacterial-type RNAP present in the mitochondrion of the heterotrophic flagellate *Reclinomonas americana* (EMBL accession no. AF007261), lending support to the endosymbiotic origin of mitochondria (Sogin, 1997). The contention that the ancestor was from the α -proteobacteria (Lang *et al.*, 1997) is supported by the close homology between the *Reclinomonas americana* insert and the corresponding insert in the β subunit of RNAP from *Rickettsia prowazekii* (Fig. 3a). A close evolutionary relationship between *Rickettsia prowazekii* and the mitochondria has previously been demonstrated by phylogenetic analysis of ribosomal proteins and

NADH dehydrogenase (Andersson *et al.*, 1998). Interestingly, and not previously reported, part of this insert appears to have been triplicated in *Reclinomonas americana* (Fig. 3b). As noted previously (Klenk *et al.*, 1999), both subunits from *Aquifex aeolicus* and *Aquifex pyrophilus* also possess this insert which, together with their position in RNAP phylogenetic trees, has led to the assertion that these species are related to the proteobacteria and do not form the deepest branch of the bacteria, as indicated by analysis based on 16S rRNA (Pitulle *et al.*, 1994), Hsp70 (Gribaldo *et al.*, 1999) and EF-G (Bocchetta *et al.*, 1995). It has been suggested that this erroneous placement of *Aquifex* species in RNAP trees is due to a distortion effect generated by the inclusion of *Mycoplasma* RNAP sequences (Bocchetta *et al.*, 2000). Even whole genome analysis of *Aquifex* has demonstrated that the phylogenetic position of this genus is ambiguous (Deckert *et al.*, 1998) and is further confused by extensive gene exchange with the archaea (Aravind *et al.*, 1998). Finally, an insert found only in the β subunits

of *Deinococcus radiodurans* (EMBL accession no. AE001944, positions 984–1027) and *Thermus aquaticus* (Y19223, positions 925–967) and containing the consensus sequence DKRE(Q/K)EVL(D/A)RA(G/E)KLG confirms the common genealogical origin of these two species proposed previously based on chemotaxonomic (Hensel *et al.*, 1986) and 16S rRNA (Weisburg *et al.*, 1989b) data.

Analysis of the RNAP β subunits divides the high- and low-G + C bacteria on the basis of amino acid content, phylogenetic analysis (Fig. 1b) and the presence of indels (Fig. 2a and b). The phylogenetic and amino acid content analyses of the species examined further divides the low-G + C bacteria into a clostridial clade and a *Bacillus*, *Listeria*, *Staphylococcus* and leuconostocs clade, with the former forming an early branch after the common Gram-positive ancestor (Fig. 1b), as described previously for other chronometers (Ahmad *et al.*, 2000; Olsen *et al.*, 1994; Gribaldo *et al.*, 1999). Although a number of inserts can be identified in one group of the low-G + C bacteria but not in the others [e.g. VHKG(D/M/I) in *Leuconostoc mesenteroides* (115–119) and *Weissella paramesenteroides* (110–114)], they also have equivalents in the Gram-negative bacteria, making their use in differentiating the low-G + C bacteria problematic. Although no indels support the previously proposed specific relationships between the Gram-positive bacteria and either mycoplasmas (Weisburg *et al.*, 1989a) or *Borrelia burgdorferi* (Eisen, 1995), there is an insert, (S/G)(S/C/A)(T/V)(L/K/F)PNRD, found in the mycobacteria (*Mycobacterium tuberculosis*, 989–996; *Mycobacterium leprae*, 991–998; *Mycobacterium smegmatis*, 980–987), *Amycolatopsis mediterranei* (980–987) and *Tropheryma whippelii* (1036–1043), that might be homologous to an insert found in the Gram-negative bacterium *Neisseria meningitidis* (NLAYPSED, 1234–1241).

ACKNOWLEDGEMENTS

This work was supported by grants from the BBSRC and the European Commission (BI02-CT94-3098).

REFERENCES

- Ahmad, S., Selvapandian, A. & Bhatnagar, R. K. (1999). A protein-based phylogenetic tree for Gram-positive bacteria derived from *hrcA*, a unique heat-shock regulatory gene. *Int J Syst Bacteriol* **49**, 1387–1394.
- Ahmad, S., Selvapandian, A. & Bhatnagar, R. K. (2000). Phylogenetic analysis of Gram-positive bacteria based on *grpE*, encoded by the *dnaK* operon. *Int J Syst Evol Microbiol* **50**, 1761–1766.
- Andersson, S. G., Zomorodipour, A., Andersson, J. O. & 7 other authors (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**, 133–140.
- Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. & Koonin, E. V. (1998). Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet* **14**, 442–444.
- Aubry-Damon, H., Soussy, C.-J. & Courvalin, P. (1998). Characterization of mutations in the *rpoB* gene that confer rifampin resistance in *Staphylococcus aureus*. *Antimicrob Agents Chemother* **42**, 2590–2594.
- Bocchetta, M., Ceccarelli, E., Creti, R., Sanangelantoni, A. M., Tiboni, O. & Cammarano, P. (1995). Arrangements and nucleotide sequence of the gene (*fus*) encoding elongation factor G (EF-G) from the hyperthermophilic bacterium *Aquifex pyrophilus*: phylogenetic depth of hyperthermophilic bacteria inferred from analysis of the EF-G/*fus* sequences. *J Mol Evol* **41**, 803–812.
- Bocchetta, M., Gribaldo, S., Sanangelantoni, A. & Cammarano, P. (2000). Phylogenetic depth of the bacterial genera *Aquifex* and *Thermotoga* inferred from analysis of ribosomal protein, elongation factor, and RNA polymerase subunit sequences. *J Mol Evol* **50**, 366–380.
- Bustard, K. & Gupta, R. S. (1997). The sequence of heat shock protein 40 (DnaJ) homologs provide evidence for a close evolutionary relationship between the *Deinococcus-Thermus* group and cyanobacteria. *J Mol Evol* **45**, 193–205.
- Collins, M. D., Love, D. N., Karjalainen, J. & 8 other authors (1994). Phylogenetic analysis of members of the genus *Porphyromonas* and description of *Porphyromonas cangingivalis* sp. nov. and *Porphyromonas cansulci* sp. nov. *Int J Syst Bacteriol* **44**, 674–679.
- Deckert, G., Warren, P. V., Gaasterland, T. & 12 other authors (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**, 353–358.
- Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* **12**, 387–395.
- Eisen, J. A. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J Mol Evol* **41**, 1105–1123.
- Enright, M., Zawadzki, P., Pickerill, P. & Dowson, C. G. (1998). Molecular evolution of rifampicin resistance in *Streptococcus pneumoniae*. *Microb Drug Resist* **4**, 65–70.
- Felsenstein, J. (1989). PHYLIP – Phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166.
- Gaurivaud, P., Laigret, F. & Bove, J.-M. (1996). Insusceptibility of members of the class *Mollicutes* to rifampin: studies of the *Spiroplasma citri* RNA polymerase β -subunit gene. *Antimicrob Agents Chemother* **40**, 858–862.
- Gribaldo, S., Lumia, V., Creti, R., de Macario, E. C., Sanangelantoni, A. & Cammarano, P. (1999). Discontinuous occurrence of the *hsp70* (*dnaK*) gene among *Archaea* and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. *J Bacteriol* **181**, 434–443.
- Gupta, R. S. (1998). Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among archaeobacteria, eubacteria, and eukaryotes. *Microbiol Mol Biol Rev* **62**, 1435–1491.
- Haas, E. S., Banta, A. B., Harris, J. K., Pace, N. R. & Brown, J. W. (1996). Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. *Nucleic Acids Res* **24**, 4775–4782.
- Hansmann, S. & Martin, W. (2000). Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol* **50**, 1655–1663.
- Harland, N. M., Leigh, J. A. & Collins, M. D. (1993). Development of gene probes for the specific identification of *Streptococcus uberis* and *Streptococcus parauberis* based upon large subunit rRNA gene sequences. *J Appl Bacteriol* **74**, 526–531.
- Hasegawa, M. & Hashimoto, T. (1993). Ribosomal RNA trees misleading? *Nature* **361**, 23.
- Heep, M., Beck, D., Bayerdörffer, E. & Lehn, N. (1999). Rifampin and rifabutin resistance mechanisms in *Helicobacter pylori*. *Antimicrob Agents Chemother* **43**, 1497–1499.
- Hensel, R., Demharter, W., Kandler, O., Kroppenstedt, R. M. & Stackebrandt, E. (1986). Chemotaxonomic and molecular-genetic studies of the genus *Thermus*: evidence for a phylogenetic relationship of *Thermus aquaticus* and *Thermus ruber* to the genus *Deinococcus*. *Int J Syst Bacteriol* **36**, 444–453.
- Huang, Y.-P. & Ito, J. (1999). DNA polymerase C of the thermophilic bacterium *Thermus aquaticus*: classification and phylogenetic analysis of the family C DNA polymerases. *J Mol Evol* **48**, 756–769.
- Jin, D. J., Cashel, M., Friedman, D. I., Nakamura, Y., Walter, W. A. & Gross, C. A. (1988). Effects of rifampicin resistant *rpoB* mutations on

- antitermination and interaction with *nusA* in *Escherichia coli*. *J Mol Biol* **204**, 247–261.
- Klenk, H. P. & Zillig, W. (1994). DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. *J Mol Evol* **38**, 420–432.
- Klenk, H. P., Meier, T.-D., Durovic, P., Schwass, V., Lottspeich, F., Dennis, P. P. & Zillig, W. (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.
- Klimpel, K. W. & Clark, V. L. (1990). The RNA polymerases of *Porphyromonas gingivalis* and *Fusobacterium nucleatum* are unrelated to the RNA polymerase of *Escherichia coli*. *J Dent Res* **69**, 1567–1572.
- Lang, B. F., Burger, G., O'Kelly, C. J., Cedergren, R., Golding, G. B., Lemieux, C., Sankoff, D., Turmel, M. & Gray, M. W. (1997). An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**, 493–497.
- Lecointre, G., Philippe, H., Van Le, H. L. & Guyader, H. (1993). Species sampling has a major impact on phylogenetic inference. *Mol Phylogenet Evol* **2**, 205–224.
- Liberal, M. H. T. & Boughton, E. (1991). Preliminary observations on *in vitro* sensitivity of *Mycoplasma* and *Acholeplasma* strains of bovine origin. *Rev Microbiol* **22**, 55–59.
- Lorenz, M. G. & Sikorski, J. (2000). The potential for intraspecific horizontal gene exchange by natural genetic transformation: sexual isolation among genomovars of *Pseudomonas stutzeri*. *Microbiology* **146**, 3081–3090.
- Ludwig, W., Neumaier, J., Klugbauer, N. & 9 other authors (1994). Phylogenetic relationships of bacteria based on comparative sequence analysis of elongation factor Tu and ATP-synthase beta-subunit genes. *Antonie Van Leeuwenhoek* **64**, 285–305.
- Maidak, B. L., Larsen, N., McCaughey, M. J., Overbeek, R., Olsen, G. J., Fogel, K., Blandy, J. & Woese, C. R. (1994). The Ribosomal Database Project. *Nucleic Acids Res* **22**, 3485–3487.
- Mollet, C., Drancourt, M. & Raoult, D. (1997). *rpoB* sequence analysis as a novel basis for bacterial identification. *Mol Microbiol* **26**, 1005–1011.
- Mollet, C., Drancourt, M. & Raoult, D. (1998). Determination of *Coxiella burnetii* *rpoB* sequence and its use for phylogenetic analysis. *Gene* **207**, 97–103.
- Morse, R., Collins, M. D., O'Hanlon, K., Wallbanks, S. & Richardson, P. T. (1996a). Analysis of the β' subunit of DNA-dependent RNA polymerase does not support the hypothesis inferred from 16S rRNA analysis that *Oenococcus oeni* (formerly *Leuconostoc oenos*) is a tachytelic (fast-evolving) bacterium. *Int J Syst Bacteriol* **46**, 1004–1009.
- Morse, R., Collins, M. D., Baldson, J. T., Wallbanks, S. & Richardson, P. T. (1996b). Nucleotide sequence of part of the *rpoC* gene encoding the β' subunit of DNA-dependent RNA polymerase from some gram-positive bacteria and comparative amino acid sequence analysis. *Syst Appl Microbiol* **19**, 150–157.
- Morse, R., O'Hanlon, K., Virji, M. & Collins, M. D. (1999). Isolation of rifampin-resistant mutants of *Listeria monocytogenes* and their characterization by *rpoB* gene sequencing, temperature sensitivity for growth, and interaction with an epithelial cell line. *J Clin Microbiol* **37**, 2913–2919.
- Naryshkina, T., Mustaev, A., Darst, S. A. & Severinov, K. (2001). The β' subunit of *Escherichia coli* RNA polymerase is not required for interaction with initiating nucleotide but is necessary for interaction with rifampicin. *J Biol Chem* **276**, 13308–13313.
- Olsen, G. J., Woese, C. R. & Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol* **176**, 1–6.
- Philippe, H., Budin, K. & Moreira, D. (1999). Horizontal transfers confuse the prokaryotic phylogeny based on the HSP70 protein family. *Mol Microbiol* **31**, 1007–1010.
- Pitulle, C., Yang, Y., Marchiani, M., Moore, E. R. B., Siefert, J. L., Aragno, M., Jurtshuk, P., Jr & Fox, G. E. (1994). Phylogenetic position of the genus *Hydrogenobacter*. *Int J Syst Bacteriol* **44**, 620–626.
- Rowland, G., Aboshkiwa, M. & Coleman, G. (1992). Comparative sequence analysis and predicted phylogeny of the DNA-dependent RNA polymerase β subunits of *Staphylococcus aureus* and other eubacteria. *Biochem Soc Trans* **21**, 40S.
- Segal, R. & Ron, E. Z. (1996). Regulation and organization of the *groE* and *dnaK* operons in eubacteria. *FEMS Microbiol Lett* **138**, 1–10.
- Shah, H. N., Gharbia, S. E. & Collins, M. D. (1997). The Gram stain: a declining synapomorphy in an emerging evolutionary tree. *Rev Med Microbiol* **8**, 103–110.
- Singer, G. A. C. & Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* **17**, 1581–1588.
- Sogin, M. L. (1997). History assignment: when was the mitochondrion founded? *Curr Opin Genet Dev* **7**, 792–799.
- Swenson, J. M., Facklam, R. R. & Thornsberry, C. (1990). Antimicrobial susceptibility of vancomycin-resistant *Leuconostoc*, *Pedococcus*, and *Lactobacillus* species. *Antimicrob Agents Chemother* **34**, 543–549.
- Tekaia, F., Lazcano, A. & Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res* **9**, 550–557.
- Telenti, A., Imboden, P., Marchesi, F., Schmidheini, T. & Bodmer, T. (1993). Direct, automated detection of rifampin-resistant *Mycobacterium tuberculosis* by polymerase chain reaction and single-strand conformation polymorphism analysis. *Antimicrob Agents Chemother* **37**, 2054–2058.
- Wang, Y. & Zhang, Z. (2000). Comparative sequence analyses reveal frequent occurrence of short segments containing an abnormally high number of non-random base variations in bacterial rRNA genes. *Microbiology* **146**, 2845–2854.
- Weisburg, W. G., Giovannoni, S. J. & Woese, C. R. (1989b). The *Deinococcus-Thermus* phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst Appl Microbiol* **11**, 128–134.
- Wichelhaus, T. A., Schafer, V., Brade, V. & Boddingtonhaus, B. (1999). Molecular characterization of *rpoB* mutations conferring cross-resistance to rifamycins on methicillin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother* **43**, 2813–2816.
- Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A* **97**, 8392–8396.