Phylogenetic analysis of Gram-positive bacteria based on grpE, encoded by the dnaK operon

Suhail Ahmad,1 Angamuthu Selvapandiyan2 and Raj K. Bhatnagar2

Author for correspondence: Suhail Ahmad. Tel: +965 531 2300 ext. 6503. Fax: +965 533 2719. e-mail: suhail.ah@hsc.kuniv.edu.kw

The dnaK operon in Gram-positive bacteria includes grpE, dnaJ and, in some members, hrcA as well. Both DnaK and DnaJ have been utilized for constructing phylogenetic relationships among various organisms. Multiple copies exist for dnaK and dnaJ genes in some bacterial genera, as opposed to a single gene copy for grpE and for hrcA, according to the currently available data. Here, we present a partial protein-based phylogenetic tree for Gram-positive bacteria, derived by using the amino acid sequence identity of GrpE; the results are compared with the phylogenetic trees generated from 5S rRNA, 16S rRNA, dnaK and dnaJ sequences. Our results indicate three main groupings: two are within low-G+C DNA Gram-positive bacteria comprising Bacillus species and Staphylococcus aureus on the one hand and Streptococcus species/Lactococcus lactis/Enterococcus faecalis/Lactobacillus sakei on the other hand; the Mycobacterium species and Streptomyces coelicolor, belonging to the high-G+C DNA Gram-positive bacteria, form the third cluster. This hierarchical arrangement is in close agreement with that obtained with 16S rRNA and DnaK sequences but not DnaJ-based phylogeny.

Keywords: grpE, dnaK operon, Gram-positive bacteria, phylogeny

The evolutionary relationships among microorganisms are being determined by comparing the sequences of rRNA genes, mainly because of their ubiquity and their conservative resistance to evolutionary changes (Olsen & Woese, 1993; Olsen et al., 1994; Stackebrandt & Goebel, 1994; Stackebrandt et al., 1997; Woese, 1994). However, the resolution power of rRNA sequences is limited among bacterial groupings such as Gram-positive bacteria that diverged at nearly the same time (Fox et al., 1992; Olsen et al., 1994; Stackebrandt et al., 1997). Gene-fusion events that produced bifunctional proteins provide the most definitive markers of evolutionary branching among bacterial groupings that diverged at nearly the same time (Ahmad & Jensen, 1988, 1989; Jensen & Ahmad, 1990). However, gene-fusion events are rare and are thus of limited use in determining bacterial phylogenies. More recently, other highly conserved molecules, e.g. heat-shock proteins (HSPs) have been used for phylogenetic analyses because of their ubiquity and their high degree of sequence conservation. These analyses have revealed a number of important differences with respect to rRNA-based phylogenies (Boorstein et al., 1994; Bustard & Gupta, 1997; Gupta, 1995, 1998).

The Gram-positive group of bacteria includes organisms that are highly pathogenic and potent toxin producers as well as non-pathogenic bacteria that are extensively used for the industrial production of antibiotics and other metabolites. A comparison of the phylogenetic trees derived by 5S rRNA sequencing (Olsen et al., 1994), 16S rRNA sequencing (de Vaux et al., 1998; Macy et al., 1996; Morse et al., 1996; Stackebrandt et al., 1997; Tsakalidou et al., 1998; Wilde et al., 1997; Zaitsev et al., 1998) and amino acid sequence comparisons for two of the HSPs encoded by the dnaK operon, DnaK (Gupta, 1998) and DnaJ (Bustard & Gupta, 1997) is shown in Fig. 1. Although exactly the same set of organisms is not shown, the discrepancies in the branching order among the trees are readily apparent. The tree based on 16S rRNA sequences depicts Bacillus stearothermophilus branching earlier than the branching of Lactococcus/Streptococcus species from other Bacillus/Staphylococcus species (Zaitsev et al., 1998). The tree based on DnaJ sequences places Mycobacterium tuberculosis closer to S. coelicolor than to Mycobacterium leprae, a result that is highly unlikely (Cole et al., 1998). However, the

Abbreviation: HSP, heat-shock protein.
The discrepancies in the branching order could result from the analysis of a small number of cistrons being misleading. In principle, the eventual comparative sequence analyses of as many carefully chosen cistrons as possible will yield slightly different trees; the greater the number of available trees, the greater will be the resolution of evolutionary branching. Thus, comparative sequence analyses of other highly conserved cistrons may help to resolve the hierarchical order.

In Gram-positive bacteria, dnaK is associated with several other heat-shock genes arranged in an operon. In Gram-positive bacteria with a high G+C DNA content, the dnaK operon consists of three genes arranged in the order dnaK–grpE–dnaJ (Buca et al., 1995; Cole et al., 1998; Jayaraman et al., 1997). However, in members of low-G+C Gram-positive bacteria, hrcA is also associated with the dnaK operon (Eaton et al., 1993; Falah & Gupta, 1997; Herbert et al., 1996; Mogk & Schumann, 1997; Narberhaus et al., 1992; Ohta et al., 1994; Wetzstein et al., 1992). In contrast to the dnaK and dnaJ genes, which are found as multiple copies, the hrcA gene and the grpE gene have each been found only as a single gene copy (Blattner et al., 1997; Cole et al., 1998; Grandvalet et al., 1998; Nimura et al., 1994; Ward-Rainey et al., 1997). GrpE is also ubiquitous and is a highly conserved protein. Thus, it was of interest to construct a phylogenetic tree, for Gram-positive bacteria, based on GrpE sequence comparisons and to compare the results obtained with the analyses derived from other HSPs encoded by the dnaK operon as well as with 5S rRNA- and 16S rRNA-derived phylogenies.

The dnaK operon from Bacillus sphaericus contains at least five protein-coding genes arranged in the order hrcA–grpE–dnaK–dnaJ–ORF35 (Ahmad et al., 1999b, EMBL accession no. Y17157). The sequence of the GrpE protein was deduced from the DNA sequence of the grpE gene and a FASTA search of EMBL, GenBank and other databases was performed using the encoded protein sequence (Altschul et al., 1990). The sequences of grpE homologues from various Gram-positive bacteria, identified in FASTA searches, were retrieved. The SWISS-PROT accession numbers of the grpE homologous sequences from the databases are as follows: B. sphaericus, O69267; Bacillus subtilis, P15874; B. stearothermophilus, Q59240; Clostridium acetobutylicum, P30726; S. aureus, P45553; L. lactis, P42369; Streptococcus mutans, O06941; L. sakei, O87776; Mycoplasma pneumoniae, P78017; Mycoplasma genitalium, P47443; Mycoplasma capricolum, P71499; M. tuberculosis, P32724; S. coelicolor, Q05562. Additionally, preliminary sequence data for the grpE sequences from Streptococcus pneumoniae, E. faecalis and Mycobacterium avium were retrieved from the Institute for Genome Research website (http://www.tigr.org). Thus, complete GrpE sequences were obtained from 16 different Gram-positive bacteria belonging to both major groupings (i.e. those with high-G+C and low-G+C DNA content).

Fig. 1. Comparison of phylogenetic relationships among Gram-positive bacteria deduced from (A) 5S rRNA sequencing (Olsen et al., 1994), (B) 16S rRNA sequencing (de Vaux et al., 1996; Morse et al., 1998; Wilde et al., 1996; Stackebrandt et al., 1997; Zaitsev et al., 1998; Nimura et al., 1994; Ward-Rainey et al., 1997). GrpE is also ubiquitous and is a highly conserved protein. Thus, it was of interest to construct a phylogenetic tree, for Gram-positive bacteria, based on GrpE sequence comparisons and to compare the results obtained with the analyses derived from other HSPs encoded by the dnaK operon as well as with 5S rRNA- and 16S rRNA-derived phylogenies.

The grpE operon from Bacillus sphaericus contains at least five protein-coding genes arranged in the order hrcA–grpE–dnaK–dnaJ–ORF35 (Ahmad et al., 1999b, EMBL accession no. Y17157). The sequence of the GrpE protein was deduced from the DNA sequence of the grpE gene and a FASTA search of EMBL, GenBank and other databases was performed using the encoded protein sequence (Altschul et al., 1990). The sequences of grpE homologues from various Gram-positive bacteria, identified in FASTA searches, were retrieved. The SWISS-PROT accession numbers of the grpE homologous sequences from the databases are as follows: B. sphaericus, O69267; Bacillus subtilis, P15874; B. stearothermophilus, Q59240; Clostridium acetobutylicum, P30726; S. aureus, P45553; L. lactis, P42369; Streptococcus mutans, O06941; L. sakei, O87776; Mycoplasma pneumoniae, P78017; Mycoplasma genitalium, P47443; Mycoplasma capricolum, P71499; M. tuberculosis, P32724; S. coelicolor, Q05562. Additionally, preliminary sequence data for the grpE sequences from Streptococcus pneumoniae, E. faecalis and Mycobacterium avium were retrieved from the Institute for Genome Research website (http://www.tigr.org). Thus, complete GrpE sequences were obtained from 16 different Gram-positive bacteria belonging to both major groupings (i.e. those with high-G+C and low-G+C DNA content).
Various grpE-encoded sequences were initially aligned using the CLUSTAL W program (http://www2.ebi.ac.uk/clustalw). Since the length of the grpE protein sequences from low-G+C Gram-positive bacteria varied significantly (174 amino acids in *S. pneumoniae* versus 221 amino acids in *B. stearothermophilus*), a pairwise comparison of GrpE sequences was carried out for the entire amino acid sequences of the GrpE protein. The alignment was performed by using the treeing algorithms contained in the PHYLIP version 3.5 phylogenetic software package. The neighbour-joining tree was deduced by using the KITSCH program contained within the software package. The robustness of tree topologies was evaluated by bootstrap analyses of the neighbour-joining data, by performing 100 resamplings (Felsenstein, 1985).

We have recently cloned and analysed the *dnaK* operon from *B. sphaericus*. The GrpE protein is encoded by the *dnaK* operon in all Gram-positive bacteria that have been studied so far (Ahmad et al., 1999a). The complete sequence of the GrpE protein is available from nearly 40 bacterial species, including 16 from the Gram-positive bacteria; so far, only a single gene copy has been reported for the *grpE* in all of these bacterial genera. This is in contrast to the multiple gene copies reported for the *dnaK* and *dnaJ* genes in several bacterial genera including Gram-positive bacteria (Blattner et al., 1997; Cole et al., 1998; Grandvalet et al., 1998; Nimura et al., 1994; Ward-Rainey et al., 1997). It is probable that the *dnaK* and/or *dnaJ* homologues in some of the bacterial genera were acquired through horizontal transfer followed by loss of the ancestral copy in some organisms. On the other hand, the presence of a single *grpE* gene across bacterial genera represents an ancestral gene copy implying a low probability of horizontal transfer. Thus, GrpE appears to be an ideal basis for the construction of phylogenetic relationships. The GrpE protein from *B. sphaericus* was compared with all of the GrpE sequences from Gram-positive bacteria and a multiple alignment of these sequences is shown in Fig. 2. The alignment is shown for the segment of sequences that is common to all these sequences. The latter sequences exhibit maximum identity among various GrpE homologues. The N-terminal residues were not included for alignment as this region varies greatly and exhibits low levels of amino acid identity among various GrpE homologues.
Table 1. Similarity matrix of grpE sequences

The pairwise amino acid identity for the entire length of various grpE homologues was calculated using the FASTA program and the PALIGN program (PCGENE).

<table>
<thead>
<tr>
<th>Organism</th>
<th>Total amino acids</th>
<th>Pairwise amino acid identity between sequences (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10 11 12 13 14 15</td>
<td></td>
</tr>
<tr>
<td>1. Bacillus subtilis</td>
<td>187</td>
<td>100</td>
</tr>
<tr>
<td>2. Bacillus stearothermophilus</td>
<td>221</td>
<td>59 100</td>
</tr>
<tr>
<td>3. Bacillus sphaericus</td>
<td>198</td>
<td>49 50 100</td>
</tr>
<tr>
<td>4. Staphylococcus aureus</td>
<td>208</td>
<td>51 40 54 100</td>
</tr>
<tr>
<td>5. Clostridium acetobutylicum</td>
<td>200</td>
<td>43 38 35 40 100</td>
</tr>
<tr>
<td>6. Lactobacillus sakei</td>
<td>197</td>
<td>42 41 42 44 36 100</td>
</tr>
<tr>
<td>7. Enterococcus faecalis</td>
<td>187</td>
<td>42 43 41 44 42 48 100</td>
</tr>
<tr>
<td>8. Lactococcus lactis</td>
<td>179</td>
<td>42 39 38 42 38 47 53 100</td>
</tr>
<tr>
<td>9. Streptococcus mutans</td>
<td>180</td>
<td>37 36 36 40 37 42 48 66 100</td>
</tr>
<tr>
<td>10. Streptococcus pneumonia</td>
<td>174</td>
<td>41 42 43 40 49 56 68 68 100</td>
</tr>
<tr>
<td>11. Mycoplasma genitalium</td>
<td>217</td>
<td>28 29 27 26 23 27 28 27 27 28 100</td>
</tr>
<tr>
<td>12. Mycoplasma pneumoniae</td>
<td>217</td>
<td>30 29 25 25 22 28 26 28 26 29 73 100</td>
</tr>
<tr>
<td>13. Mycobacterium tuberculosis</td>
<td>235</td>
<td>28 20 28 27 26 30 24 28 26 28 18 18 100</td>
</tr>
<tr>
<td>14. Mycobacterium avium</td>
<td>227</td>
<td>31 31 31 29 29 26 29 28 29 30 19 19 77 100</td>
</tr>
<tr>
<td>15. Streptomyces coelicolor</td>
<td>225</td>
<td>29 27 34 29 28 26 28 29 28 30 18 18 37 36 100</td>
</tr>
</tbody>
</table>

bacteria, using the FASTA program and the PALIGN program. Although the multiple alignment shown in Fig. 2 was derived for the common segment in all grpE sequences, pairwise comparisons were performed for the entire length of various GrpE homologues. This analysis indicated that the amino acid identity between various homologues varied between 18 and 77% over the entire length of these proteins (Table 1). The grpE sequences from the two Mycoplasma species yielded significantly lower amino acid identity with the grpE homologues from Gram-positive bacteria with high G+C DNA content (the two Mycobacterium species and S. coelicolor) (Table 1). This is because of low levels of amino acid identity near the N- and C-terminal ends of these grpE sequences (data not shown).

The amino acid sequence alignment of the grpE homologues shown in Fig. 2 was used to examine the phylogenetic relationships between various Gram-positive bacteria. A neighbour-joining consensus tree based on GrpE sequences is shown in Fig. 3. The observed bootstrap scores for all of the nodes are indicated. The branching order, rather than the actual distances on the tree, is shown. The tree showed that the Gram-positive bacteria form three groupings, i.e. two within low-G+C DNA Gram-positive bacteria and a third comprising those with high-G+C DNA content. All of the members of the low-G+C group of Gram-positive bacteria, except the two Mycoplasma species, share an immediate common ancestor. These Mycoplasma species carry the smallest genome of any free-living organism (Fraser et al., 1995; Himmelreich et al., 1996) and the grpE sequences may have diverged much more in these organisms as a result of reduction in the genome size. The three Bacillus species used in this study and S. aureus share an immediate common ancestor, demonstrating their close evolutionary re-
The two Streptococcus species/L. lactis/E. faecalis/L. sakei formed another grouping within the low-G+C Gram-positive bacteria, whereas *M. tuberculosis*, *M. avium* and *S. coelicolor* (Gram-positive bacteria with high-G+C DNA content) formed the third cluster. These hierarchical arrangements are supported by the 5S rRNA-, 16S rRNA- and DnaK-based phylogenies (Gupta, 1998; de Vaux *et al*., 1994). However, it should be noted that the phylogenetic trees based on HrcA sequences places them in the *Bacillus* species/S. aureus cluster (Ahmad *et al*., 1999a). The tree based on DnaJ sequences shows *Clostridium* species branching at the deepest level, even earlier than the branching of Gram-positive bacteria with high-G+C DNA content from those with low-G+C DNA (Bustard & Gupta, 1997).

The data presented here show that the phylogenetic positioning of Gram-positive bacteria based on *grpE* sequences is generally in agreement with 16S rRNA- and DnaK-based phylogenies but not with DnaJ-based phylogeny. The discrepancies in the branching orders noted above could result from different mutation rates for different genes in these organisms following their divergence from the common ancestor. Alternatively, it is also possible that the gene under study (e.g. *dnaJ*) was acquired through horizontal transfer in one lineage at an earlier stage of divergence. It has been reported that some bacterial genera including Gram-positive bacteria contain more than one gene copy of *dnaK* and *dnaJ* genes (Blattner *et al*., 1997; Cole *et al*., 1998; Grandvalet *et al*., 1998; Ward-Rainey *et al*., 1997). The extent of this duplication and the mechanism by which the two gene copies were acquired is not known at present. It is probable that one of the gene copies was acquired through horizontal transfer in these genera. If so, this may explain some of the varying results obtained with *dnaJ*-based phylogenies.

**Acknowledgements**

We thank Professor Roy A. Jensen for his suggestions and for a critical reading of the manuscript. Preliminary sequence data for the *grpE* sequences from *S. pneumoniae*, *E. faecalis* and *M. avium* were retrieved from the Institute for Genome Research website (http://www.tigr.org). The sequencing of *E. faecalis*, *M. avium* and *S. pneumoniae* was accomplished with support from the Institute of Genome Research, Rockville, MD, USA.

**References**


