NOTE

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

¹ Department of Microbiology, Faculty of Medicine, Kuwait University, PO Box 24923, Safat 13110, Kuwait

² International Centre for Genetic Engineering and Biotechnology, Aruna A. Ali Marg, PO Box 10504, New Delhi-110067, India

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 micro-organisms 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.

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.*, 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), (C) the amino acid sequence identity of DnaK proteins (Gupta, 1998) and (D) DnaJ protein sequences (Bustard & Gupta, 1997). The branching order, rather than the actual distances on the trees, is shown.

results derived from the analysis of cistrons other than rRNA genes should be interpreted with caution as hierarchical arrangements deduced from rRNA genes are based on complete sequencing of these genes from several thousand species and subspecies as opposed to the 20 up to a few hundred sequence entries for other conserved cistrons.

The discrepancies in the branching order could result either from different mutation rates for these genes in various organisms following their divergence or from horizontal transfer of the gene at an earlier point of divergence. Alternatively, the gene under study may have undergone duplication followed by a change of

function due to mutations in some lineages. Thus, a branching order deduced from the analysis of a small number of cistrons could be 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* (Bucca *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; Herbolt *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).

Bacsu	ELLQNQINELQGLLEKEN	KLLRVQADFN YKR	RSRLMEASQKYRSQN	84
Bacst	AAANAKNCRTRS-D-RGNRY-LY	FR	-T-Q--AE--A-S	
Bacsp	QYEAQVA---AK-DDD-E	NRH--IR---D-MR-	-QQ-DG--AE--A-S	
Staa	DPKDKQ-----Q-AD-N-E	-Y--LY-E-----	-IQK-N-INKT-QA-R	
Strmu	TSENKELQ-ALERA-DF--	-Y-AH-EMPKTSV	ALMKSQKVC-R---D	
Strpn	TPEKSELDLANERAD-F--	-Y-AH-EMQ- IQ-	-ANE-RQNL-R-----	
Lacla	-EEI-ELD-A-K-AT-W-	-F--S-EMQ- VQ-	-GNE-RQLV-----D	
Entfa	-VEATE-EN-KAE-S-M-D	-F-AR-EIA- MSN	-NKN-R-LVR-----D	
Labsa	-EADDKLA---KKYDAM-D	SF--S-EMQ- IQM	-NQR-QANLL-DG-S	
Cloab	LK-KDENEK-KNE-DAKD	R--LS-EY--RN	-TAK-K-GIYTDACSD	
Mygge	NE--QK-PL--KAN--FAL	-FE-M-REAQ-QIAKLDELNLKMKK-L-QAK--AIAK		
Myppn	NE--QK-PN--KVN--FRL	-VEKI-EAQKKIQEKVAELTIK-KE-L-NAK--VIEK		
Myctu	GTAADAHAHTEDKVA-LTA	D-Q-----A-RK	-ALRDQ--AADRAKAS	
Mycav	AATAESGGAADKVA-LTA	D-Q-----A-RK	-ALRDQ--AADRAKAS	
Stmco	AG-VA-LDQVTA-N-RTA	D-Q-L--EYQ- -R-	-VERDRV-VKEVAVA-	

Box 1				
Bacsu	IVTLLPALDSFERALQVEADNEQTSKLLQG	MEMVHRQLVREALKKEGVEAIEAV	GQE	141
Bacst	L-S-----N-----KI-TE--A-I--	-----Y-SVLD-----	-KP	
Bacsp	LLS-----V--N-----TTS-E-A-I-K-	I--Y-S-L-TVF--LQV-K-E	-EQ	
Staa	VL-I--I--NI-----I-G-D-TF--QK-	VQ--ES-IN--DN-L-V-KTE	-EA	
Strmu	LRKA-----S-NL-----A-G	L-DVVKK- L--QES-IQ--E--EV-LE	N	
Strpn	LAKI-----S-NL-----A-G	L-DVVKK- L-G--QES-IH--E--I-E-A-D	-	
Lacla	LAKI-----S-NL-----A-G	L-DVVKK- L--QES-IH--E--I-E-A-D	-	
Entfa	LKGKI-----SI-NL-----M-L-VSD-	GE-KK- IS--LESITV--E--I-E-P-M	-ET	
Labsa	LAK-V--V--NL-----A-A-	TDE--KK- VQ-TYDHMKH--EDH--KE--Q	-A	
Cloab	VINEM--T--NL-----AST-G	SAEDIKK- V--VK-FKNS-S-L-E-PSE	-K	
Mygge	TLDQP-NII-Q-I--SYAQKDP-V-NYTT-	FN--LDAFSRW-EAN--TK-KIEP-M		
Myppn	SIDQP-III-Q-I--SYAQKDP-V-NYTT-	FN--LDAFSRW-EGF--TK-AIEP-AQ		
Myctu	V-SQ--GV--DL--RKHG	H-ES-PLKS-ADK-DS--TGL-LV-FGE	-E	
Mycav	V-NQ--GV--DL--RKHG	D-ES-PLKS-ADK-ES--TGL-LT-FGEE	-E	
Stmco	LSH-----V--DVG--REHG	E-VG- FKS-AES-ETTVA-L-LQFGKE	-EP	

Box 2				
Bacsu	FDPNLHQAVMVA	EDENYGSNIVVEM	QKGYKLKDRVIRFSMVKNQ	187
Bacst	---H-----V--S-EP-T-----	-----A-----S		
Bacsp	---I-----E-Q-SEKETGV-IR-L	---I-----L--T-S--E		
Staa	---I-----V-D-DNPFDE-GEITQ-L	-----L-----		
Strmu	---A-----M-QTLDA-DDHPADSIAQVH	---Q-HE-LL--A--V-YN		
Strpn	---H-Y-M-QTLPA-DEHPVDITIAQVF	---H-IL--A--V-A		
Lacla	---H-----M-QTVFA-DEHPADSI-QVF	---Q-HE-LL--A--V-A		
Entfa	---QTVFAS-ETPADTI--VI	---Q--L--I-A		
Labsa	---TI-----QTVAV-GDQKADT--RVS	---Y--Y--L--A--V-A		
Cloab	---N-----H-----G-E-E--VL	---RG-K--L--H--AN		
Mygge	---EKIMS-LEIVDSN	LAK-K--RVSKS--Y-K--FAS-F-SK		
Myppn	---EKVMA-LEIVPS-	QPA-T--KVSKS--H-K--FAS-V-S		
Myctu	---V-E--QHEGSGGSGSKPVIGTV	---Q-Q-GEQ-L-HAL-G-VD		
Mycav	---V-E--QHEGSGG	SKPVIGTV-RQ--G-Q-L-HAL-G-VD		
Stmco	---TI-E-L-HSYAP	DVTETTC-AIL-P--RIGE-T--AR-A-AE		

Fig. 2. Multiple alignment of *grpE*-encoded sequences used in this study from various Gram-positive bacteria. The complete names of the species are: Bacsu, *Bacillus subtilis*; Bacst, *Bacillus stearothermophilus*; Bacsp, *Bacillus sphaericus*; Cloac, *Clostridium acetobutylicum*; Staa, *Staphylococcus aureus*; Lacla, *Lactococcus lactis*; Mygge, *Mycobacterium genitalium*; Myppn, *Mycobacterium pneumoniae*; Strmu, *Streptococcus mutans*; Strpn, *Streptococcus pneumoniae*; Labsa, *Lactobacillus sakei*; Entfa, *Enterococcus faecalis*; Myctu, *Mycobacterium tuberculosis*; Mycav, *Mycobacterium avium* and Stmco, *Streptomyces coelicolor*. The alignment is shown for the common segments that are present in all *grpE* homologues and correspond to amino acid residues 36–187 of the *B. subtilis* *grpE* protein (Narberhaus *et al.*, 1992). Amino acids are listed in the standard one-letter code and residues identical to those of *B. subtilis* are indicated by dashes. Gaps (indicated by blank spaces) were introduced to obtain a maximum fit. The three highly conserved domains are labelled as Box 1, Box 2 and Box 3, respectively.

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*) and little sequence identity was observed for the N-terminal amino acid residues, the alignment is shown for the common segments that are present in all *grpE* sequences. This segment corresponded to amino acid residues 36–187 of the *B. subtilis* GrpE protein sequence (Narberhaus *et al.*, 1992). The alignment was inspected for any obvious misalignments. However, the amino acid sequence identity between pairs of proteins was calculated for the entire

length of all *grpE* proteins by using the FASTA program and the PALIGN program (PCGENE) (Altschul *et al.*, 1990). The phylogenetic analyses were carried out with the same amino acid sequences that were used for multiple alignment and which are common to all the *grpE* homologues. The N-terminal segments of GrpE sequences for which a homologous character cannot be postulated were excluded from phylogenetic analysis. The phylogenetic relationships were determined 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 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 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 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.

Among all the GrpE sequences, three conserved regions have been identified and are labelled as boxes 1, 2 and 3, respectively. The C-terminal 50 amino acids of GrpE protein sequences exhibit maximum identity and contain two of the three conserved regions (Fig. 2, boxes 2 and 3). This is in contrast to amino acid sequence identities observed with two other proteins encoded by the *dnaK* operon; i.e. the DnaJ (Bustard & Gupta, 1997) and HrcA (Ahmad *et al.*, 1999a) sequences. The latter sequences exhibit maximum identity at the N-terminal ends.

A pairwise comparison of GrpE sequences was carried out from representative species of Gram-positive

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).

Organism	Total amino acids	Pairwise amino acid identity between sequences (%)														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. <i>Bacillus subtilis</i>	187	100														
2. <i>Bacillus stearothermophilus</i>	221	59	100													
3. <i>Bacillus sphaericus</i>	198	49	50	100												
4. <i>Staphylococcus aureus</i>	208	51	40	45	100											
5. <i>Clostridium acetobutylicum</i>	200	43	38	35	40	100										
6. <i>Lactobacillus sakei</i>	197	42	41	42	44	36	100									
7. <i>Enterococcus faecalis</i>	187	42	43	41	44	42	48	100								
8. <i>Lactococcus lactis</i>	179	42	39	38	42	38	47	53	100							
9. <i>Streptococcus mutans</i>	180	37	36	36	40	37	42	48	66	100						
10. <i>Streptococcus pneumoniae</i>	174	41	42	41	43	40	49	56	68	68	100					
11. <i>Mycoplasma genitalium</i>	217	28	29	27	26	23	27	28	27	27	28	100				
12. <i>Mycoplasma pneumoniae</i>	217	30	29	25	25	22	28	26	28	26	29	73	100			
13. <i>Mycobacterium tuberculosis</i>	235	28	20	28	27	26	30	24	28	26	28	18	18	100		
14. <i>Mycobacterium avium</i>	227	31	31	31	29	29	26	29	28	29	30	19	19	77	100	
15. <i>Streptomyces coelicolor</i>	225	29	27	34	29	28	26	28	29	28	30	18	18	37	36	100

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

Fig. 3. Consensus bootstrap neighbour-joining phylogenetic tree for Gram-positive bacteria based on GrpE sequences. The numbers on the nodes indicate how often (no. of times, %) the species to the right grouped together in 100 bootstrap samples. The phylogenetic analysis was carried out on the alignment of the common segments of all *grpE* sequences as shown in Fig. 2. The branching order, rather than the actual distances on the tree, is shown.

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-

lationship. This is also evidenced by the analyses of other highly conserved cistrons such as 5S rRNA and DnaK (Gupta, 1998; Olsen *et al.*, 1994). However, in the phylogenetic tree derived from 16S rRNA sequences *B. stearothermophilus* branches from other *Bacillus* species earlier than does *S. aureus* and is in a separate cluster (Zaitsev *et al.*, 1998). The tree topologies obtained with different markers should be interpreted with caution as some of these markers may contain very little sequence information (e.g. 5S rRNA) compared with larger cistrons (e.g. 16S rRNA, *dnaK*).

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.*, 1998; Olsen *et al.*, 1994; Tsakalidou *et al.*, 1998; Zaitsev *et al.*, 1998) but not by DnaJ-based phylogeny (Bustard & Gupta, 1997). The phylogenetic positioning of *C. acetobutylicum* and other *Clostridium* species is different in different trees. The trees based on GrpE (Fig. 3), 5S rRNA (Olsen *et al.*, 1994), 16S rRNA (Macy *et al.*, 1996; Wilde *et al.*, 1997) and DnaK (Gupta, 1998) sequences depict *Clostridium* species branching first from the common ancestor before the branching of other low-G+C Gram-positive bacteria, while the tree 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). However, it should be noted that the phylogenetic trees derived from *grpE*, *hrcA*, *dnaJ* and *dnaK* cistrons are deduced from a relatively small data set (gene sequences from approx. 20–100 organisms) compared to 16S rRNA-derived phylogenies obtained by using sequence data from thousands of organisms.

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

- Ahmad, S. & Jensen, R. A. (1988). The evolutionary history of two bifunctional proteins that emerged in the purple bacteria, *Trends Biochem Sci* **11**, 108–112.
- Ahmad, S. & Jensen, R. A. (1989). Utility of a bifunctional tryptophan pathway enzyme for the classification of the herbicola-agglomerans complex of bacteria, *Int J Syst Bacteriol* **39**, 100–104.
- Ahmad, S., Selvapandiyan, A. & Bhatnagar, R. K. (1999a). 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., Selvapandiyan, A., Gasbarri, M. & Bhatnagar, R. K. (1999b). Molecular cloning of the *dnaK* gene region from *Bacillus sphaericus* in the context of genomic comparisons, *Microb Comp Genom* **4**, 47–58.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool, *J Mol Biol* **215**, 403–410.
- Blattner, F. R., Plunkett, G., III, Bloch, C. A. & 14 other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462.
- Boorstein, W. R., Ziegelhoffer, T. & Craig, E. A. (1994). Molecular evolution of the hsp70 multigene family, *J Mol Evol* **38**, 1–17.
- Bucca, M., Ferina, G., Puglia, A. N. & Smith, C. P. (1995). The *dnaK* operon of *Streptomyces coelicolor* encodes a novel heat-shock protein which binds to the promoter region of the operon, *Mol Microbiol* **17**, 663–674.
- Bustard, K. & Gupta, R. S. (1997). The sequences of heat shock protein 40 (DnaJ) homologs provide evidence for a close evolutionary relationship between the *Deinococcus*–*Thermus* group and cyanobacteria, *J Mol Evol* **45**, 193–205.
- Cole, S. T., Brosch, R., Parkhill, J. & 39 other authors (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.
- Eaton, T., Shearman, C. & Gasson, M. (1993). Cloning and sequence analysis of the *dnaK* gene region of *Lactococcus lactis* subsp. *lactis*, *J Gen Microbiol* **139**, 3253–3264.
- Falah, M. & Gupta, R. S. (1997). Phylogenetic analysis of mycoplasmas based on hsp70 sequences: cloning of the *dnaK* (hsp70) gene region of *Mycoplasma capricolum*, *Int J Syst Bacteriol* **47**, 38–45.
- Felsenstein, J. (1985). Confidence limits in phylogenies: an approach using the bootstrap, *Evolution* **39**, 783–791.

- Fox, G. E., Wisotzkey, J. D. & Jurtshuk, P., Jr (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity, *Int J Syst Bacteriol* **42**, 166–170.
- Fraser, C. M., Gocayne, J. D., White, O. & 26 other authors (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.
- Grandvalet, C., Rapoport, G. & Mazodier, P. (1998). *hrcA*, encoding the repressor of the *groEL* genes in *Streptomyces albus* G, is associated with a second *dnaJ* gene, *J Bacteriol* **180**, 5129–5134.
- Gupta, R. S. (1995). Phylogenetic analysis of the 90 kD heat shock family of protein sequences and an examination of the relationships between animals, plants and fungi species, *Mol Biol Evol* **12**, 1063–1073.
- 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.
- Herbort, M., Schon, U., Angermann, K., Lang, J. & Schumann, W. (1996). Cloning and sequencing of the *dnaK* operon of *Bacillus stearothermophilus*, *Gene* **170**, 81–84.
- Himmelreich, R., Hilbert, H., Plagens, H., Pirkel, E., Li, B.-C. & Hermann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*, *Nucleic Acids Res* **24**, 4420–4449.
- Jayaraman, G. C., Penders, J. E. & Burne, R. A. (1997). Transcriptonal analysis of the *Streptococcus mutans* *hrcA*, *grpE* and *dnaK* genes and regulation of expression in response to heat shock and environmental acidification, *Mol Microbiol* **25**, 329–341.
- Jensen, R. A. & Ahmad, S. (1990). Nested gene fusions as markers of phylogenetic branch points in prokaryotes, *Trends Ecol Evol* **5**, 219–224.
- Macy, J. M., Nunan, K., Hagen, K. D., Dixon, D. R., Harbour, P. J., Cahill, M. & Sly, L. I. (1996). *Chrysiogenes arsenatis* gen. nov., sp. nov., a new arsenate-respiring bacterium isolated from gold mine wastewater, *Int J Syst Bacteriol* **46**, 1153–1157.
- Mogk, A. & Schumann, W. (1997). Cloning and sequencing of the *hrcA* gene of *Bacillus stearothermophilus*, *Gene* **194**, 133–136.
- Morse, R., Collins, M. D., O'Hanlon, K., Wallbanks, S. & Richardson, P. T. (1996). 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.
- Narberhaus, F., Giebler, K. & Bahl, H. (1992). Molecular characterization of the *dnaK* gene region of *Clostridium acetobutylicum*, including *grpE*, *dnaJ* and a new heat shock gene, *J Bacteriol* **174**, 3290–3299.
- Nimura, K., Yoshikawa, H. & Takahashi, T. (1994). Identification of *dnaK* multigene family in *Synechococcus* sp. PCC 7942, *Biochem Biophys Res Commun* **201**, 466–471.
- Ohta, T., Saito, K., Kuroda, M., Honda, K., Hirata, H. & Hayashi, H. (1994). Molecular cloning of two new heat shock genes related to the *hsp70* genes in *Staphylococcus aureus*, *J Bacteriol* **176**, 4779–4783.
- Olsen, G. J. & Woese, C. R. (1993). Ribosomal RNA: a key to phylogeny, *FASEB J* **7**, 113–123.
- Olsen, G. J., Woese, C. R. & Overbeek, R. (1994). The winds of (evolutionary) change: breathing new life into Microbiology, *J Bacteriol* **176**, 1–6.
- Stackebrandt, E. & Goebel, B. M. (1994). A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology, *Int J Syst Bacteriol* **44**, 846–849.
- Stackebrandt, E., Rainey, F. A. & Ward-Rainey, N. L. (1997). Proposal for a new hierarchic classification system, *Actinobacteria* classis nov, *Int J Syst Bacteriol* **47**, 479–491.
- Tsakalidou, E., Zoidou, E., Pot, B., Wassill, L., Ludwig, W., Devriese, L. A., Kalantzopoulos, G., Schleifer, K. H. & Kersters, K. (1998). Identification of streptococci from Greek kasseri cheese and description of *Streptococcus macedonicus* sp. nov, *Int J Syst Bacteriol* **48**, 519–527.
- de Vaux, A., Laguerre, G., Diviès, C. & Prévost, H. (1998). *Enterococcus asini* sp. nov. isolated from the caecum of donkeys (*Equus asinus*), *Int J Syst Bacteriol* **48**, 383–387.
- Ward-Rainey, N., Rainey, F. A. & Stackebrandt, E. (1997). The presence of a *dnaK* (HSP70) multigene family in members of the orders *Planctomycetales* and *Verrucomicrobiales*, *J Bacteriol* **179**, 6360–6366.
- Wetzstein, M., Volker, U., Dedio, J., Loban, S., Zuber, U., Schiesswohl, M., Herget, C., Hecker, M. & Schumann, W. (1992). Cloning, sequencing and molecular analysis of the *dnaK* locus from *Bacillus subtilis*, *J Bacteriol* **174**, 3300–3310.
- Wilde, E., Collins, M. D. & Hippe, H. (1997). *Clostridium pascui* sp. nov., a new glutamate-fermenting sporeformer from a pasture in Pakistan, *Int J Syst Bacteriol* **47**, 164–170.
- Woese, C. R. (1994). There must be a prokaryote somewhere: Microbiology's search for itself, *Microbiol Rev* **58**, 1–9.
- Zaitsev, G. M., Tsitko, I. V., Rainey, F. A., Trotsenko, Y. A., Uotila, J. S., Stackebrandt, E. & Salkinoja-Salonen, M. S. (1998). New aerobic ammonium-dependent obligately oxalotrophic bacteria: description of *Ammoniphilus oxalaticus* gen. nov., sp. nov. and *Ammoniphilus oxalivorans* gen. nov., sp. nov, *Int J Syst Bacteriol* **48**, 151–163.