Application of a *recN* sequence similarity analysis to the identification of species within the bacterial genus *Geobacillus*

Daniel R. Zeigler

Bacillus Genetic Stock Center, Department of Biochemistry, The Ohio State University, Columbus, OH 43210, USA

Correspondence
Daniel R. Zeigler
zeigler.1@osu.edu

Full-length *recN* and 16S rRNA gene sequences were determined for a collection of 68 strains from the thermophilic Gram-positive genus *Geobacillus*, members of which have been isolated from geographically and ecologically diverse locations. Phylogenetic treeing methods clustered the isolates into nine sequence similarity groups, regardless of which gene was used for analysis. Several of these groups corresponded unambiguously to known *Geobacillus* species, whereas others contained two or more type strains from species with validly published names, highlighting a need for a re-assessment of the taxonomy for this genus. For taxonomic analysis of bacteria related at a genus, species or subspecies level, *recN* sequence comparisons had a resolving power nearly an order or magnitude greater than 16S rRNA gene comparisons. Mutational saturation rendered *recN* comparisons much less powerful than 16S rRNA gene comparisons for analysis of higher taxa, however. Analysis of *recN* sequences should prove a powerful tool for assigning strains to species within *Geobacillus*, and perhaps within other genera as well.

INTRODUCTION

Bacterial systematics rests on the concept of grouping bacteria based on similarities in genome content and in observable traits (Gürtler & Mayall, 2001). For measuring genome similarity, DNA–DNA hybridization studies have been regarded as the 'gold standard' (Wayne *et al.*, 1987), but it can be difficult to reproduce hybridization values between laboratories with the necessary precision. Furthermore, hybridization methods often require specialized equipment or the use of radioactive labels. Consequently, systematists have come to rely increasingly on comparison of DNA sequences, especially 16S rRNA gene sequences, as a supplementary or alternative approach (Stackebrandt & Goebel, 1994). High-throughput facilities have made DNA sequencing rapid, reproducible and inexpensive. Public databases and sophisticated analysis software are freely available. Recently, an ad hoc committee for re-evaluating the species definition called on systematists to determine whether sequencing a set of genes could yield results congruent with DNA hybridization methods, with a view towards identifying even single genes that could be useful for assigning isolates to species (Stackebrandt *et al.*, 2002).

Recently, Zeigler (2003) identified over 30 genes that met specific criteria: (1) wide distribution among bacteria; (2) uniqueness within each genome; (3) phylogenetically informative size; and (4) sequence divergence that mirrors whole genome divergence among related species. One strong candidate as a genome similarity predictor was *recN*. For the species studied, genome identity scores predicted by *recN* analysis differed from those measured directly in genomic alignments by an average of only 4·4 %.

In the present study, sequences of both the 16S rRNA gene and *recN* were determined for a group of 68 isolates from the genus *Geobacillus* (Nazina *et al.*, 2001). The striking congruence of phylogenetic trees constructed with these sequences suggests that the two genes have experienced similar histories within the genus, and that horizontal gene transfer has not disrupted their relationship. For grouping closely related organisms, *recN* was clearly superior to the 16S rRNA gene, with nearly an order of magnitude greater resolving power at the species–subspecies level. Thus, *recN* seems to satisfy the requirements of Stackebrandt *et al.* (2002) as a useful tool for assigning isolates to species within this genus.

Published online ahead of print on 7 January 2005 as DOI 10.1099/ijs.0.63452-0.

The GenBank/EMBL/DDBJ accession number for the 16S rRNA gene sequences described in this study are AY297092 and AY608927–AY608993, inclusive; those for the *recN* sequences are AY434725 and AY608994–AY609060, inclusive.

A table giving details for the bacterial strains used in this study and a figure showing the locations of primers used for the *recN* sequencing project are available as supplementary material in IJSEM Online.

METHODS

Isolation and maintenance of *Geobacillus* strains. Novel *Geobacillus* isolates were obtained from environmental samples by mixing 0·5–2·5 g soil or other environmental samples in 5 ml sterile distilled water and vortexing the suspension thoroughly. After allowing large particles to settle for 30 min, the suspension was diluted serially in sterile water, and 0·1 ml aliquots were spread onto tryptose blood agar base (TBAB) plates. Plates were incubated at 63 °C for 18 h. Individual colonies were streak-purified on fresh TBAB at 63 °C before storage. All environmental samples were collected either from the Sangre de Cristo Mountains area of New Mexico during August 1999 or around the Ohio State University campus in Columbus, Ohio, USA during the spring of 2003. The entire *Geobacillus* collection, including the novel isolates, is listed in the Supplementary Table available in IJSEM Online.

DNA sequencing. Each isolate was grown overnight at 60 °C with vigorous aeration in 1 litre shake flasks containing 50 ml liquid medium – Luria broth, brain heart infusion or TBAB-B (10·0 g tryptose, 3·0 g beef extract and 5·0 g NaCl per litre of water). Genomic DNA was isolated from the culture by using the Qiagen Genomic-tip 500/G kit according to the manufacturer's instructions, except that the cleared lysate was vortexed at high speed for 30 s prior to loading on the binding column. DNA sequences were obtained directly from genomic DNA samples, without amplification or subcloning, using custom primers designed for *recN* or the 16S rRNA gene. For 16S rRNA gene sequencing, primers were pA and pD(R) (Edwards *et al.*, 1989), 765r and 1495r (Lu *et al.*, 2001), 16F358, 16F926 and 16R1093 (Coenye *et al.*, 1999), and 16F1074, which is the reverse complement of 16R1093. A complete list of the primers used for *recN* sequencing is available with the Supplementary Figure in IJSEM Online. DNA sequences were determined on an automated 3730 DNA Analyser (Applied Biosystems), using BigDye terminator cycle sequencing, following the manufacturer's specifications for genomic DNA.

Sequence analysis. DNA sequences were assembled with SeqManII (DNASTAR, Madison, WI, USA). Multiple alignments and distance matrices were constructed by using CLUSTAL W (Thompson *et al.*, 1994). DNA alignments were hand-corrected to ensure that they were consistent with predicted amino acid alignments for each gene product. Phylogenetic trees were constructed by using the NEIGHBOR application of the PHYLIP software package (Felsenstein, 1989) and visualized by using TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html). Unrooted parsimony analysis was conducted on CLUSTAL W-generated multiple alignments with the PHYLIP DNAPARS application. Statistical analysis was performed with Sigma Plot.

RESULTS AND DISCUSSION

Recently, Zeigler (2003) proposed that *recN* sequence comparisons could accurately measure genome similarities for a wide range of bacterial taxa. Members of the Gram-positive thermophilic genus *Geobacillus* (Nazina *et al.*, 2001) were chosen as a test set for validating *recN* analysis as a taxonomic tool. A total of 48 *Geobacillus* isolates was obtained from public culture collections in Germany, Japan and the United States, as well as from individual researchers in Australia, Italy, Turkey, Japan and the United States. Bacteria in this collection had been isolated from a wide variety of sources, including geothermal waters and soils, spoiled foods, composted organic matter and temperate soils. Included in the collection were the type strains or other representatives of each of the *Geobacillus* species that had validly published names at the beginning of the study. The collection was supplemented with 20 novel isolates obtained from soil and other environmental samples collected from two locations with mesic temperature regimes: one an arid, rural region of the southwestern United States and the other a well-watered, urban area in the mid-western United States. Altogether, the strains in this collection should include a broad sampling of the genus *Geobacillus*, presenting the kind of taxonomic challenges typically faced with bacterial isolates assembled during a moderately sized discovery programme. The strains in the collection, together with GenBank/EMBL/DDBJ accession numbers for sequence data, are listed in the Supplementary Table in IJSEM Online.

For each of the 68 strains, high-quality, full-length DNA sequences were determined for both *recN* and the 16S rRNA gene. Because 16S rRNA gene sequencing has become a standard procedure in characterizing a new bacterial isolate, selection of suitable primers was a simple matter. Sequencing the *recN* gene, which is much less highly conserved than the 16S rRNA gene (Zeigler, 2003), was a considerable challenge. Primer selection was greatly aided by aligning genomic DNA sequences for several members of the family *Bacillaceae* (not shown) with the unfinished genome sequence of *Geobacillus stearothermophilus* strain 10 (=BGSC 9A21) [B. Roe, *Bacillus* (*Geobacillus*) *stearothermophilus* Genome Sequencing Project, http://www.genome.ou.edu/bstearo.html]. The gene order *spoIVB–recN–ahrC* is very highly conserved among the *Bacillaceae* and *Clostridiaceae*; the order *recN–ahrC* is conserved even more widely within the phylum *Firmicutes* (unpublished data). Because the sequences of *spoIVB* and *ahrC* are more highly conserved than *recN*, it was possible to design primers flanking *recN* that allowed for direct sequencing of genomic DNA in the *Geobacillus* isolates. Partial *recN* sequences provided enough data to allow for a 'primer walking' strategy to sequence the remainder of the gene in each strain in the collection. The *recN* sequencing primers are detailed in the Supplementary Figure in IJSEM Online.

Comparison of *recN* and 16S rRNA gene phylogenies for *Geobacillus*

Phylogenies constructed with the 16S rRNA and *recN* gene sequences are remarkably similar for the strains in the *Geobacillus* collection (Fig. 1). Each phylogram clusters the 68 strains into the same sequence similarity groups. The main difference between the phylograms is in branch length. In particular, the branches separating the nine sequence similarity groups are especially elongated in the *recN* tree relative to the 16S rRNA tree. Bootstrap support is strong for the groups in both trees, but is nearly unanimous for the *recN* phylogram.

A more quantitative demonstration of the similarities between *recN* and 16S rRNA sequence analysis for this strain set can be obtained by plotting for each pair of strains the frequency of identical residues in *recN* sequence

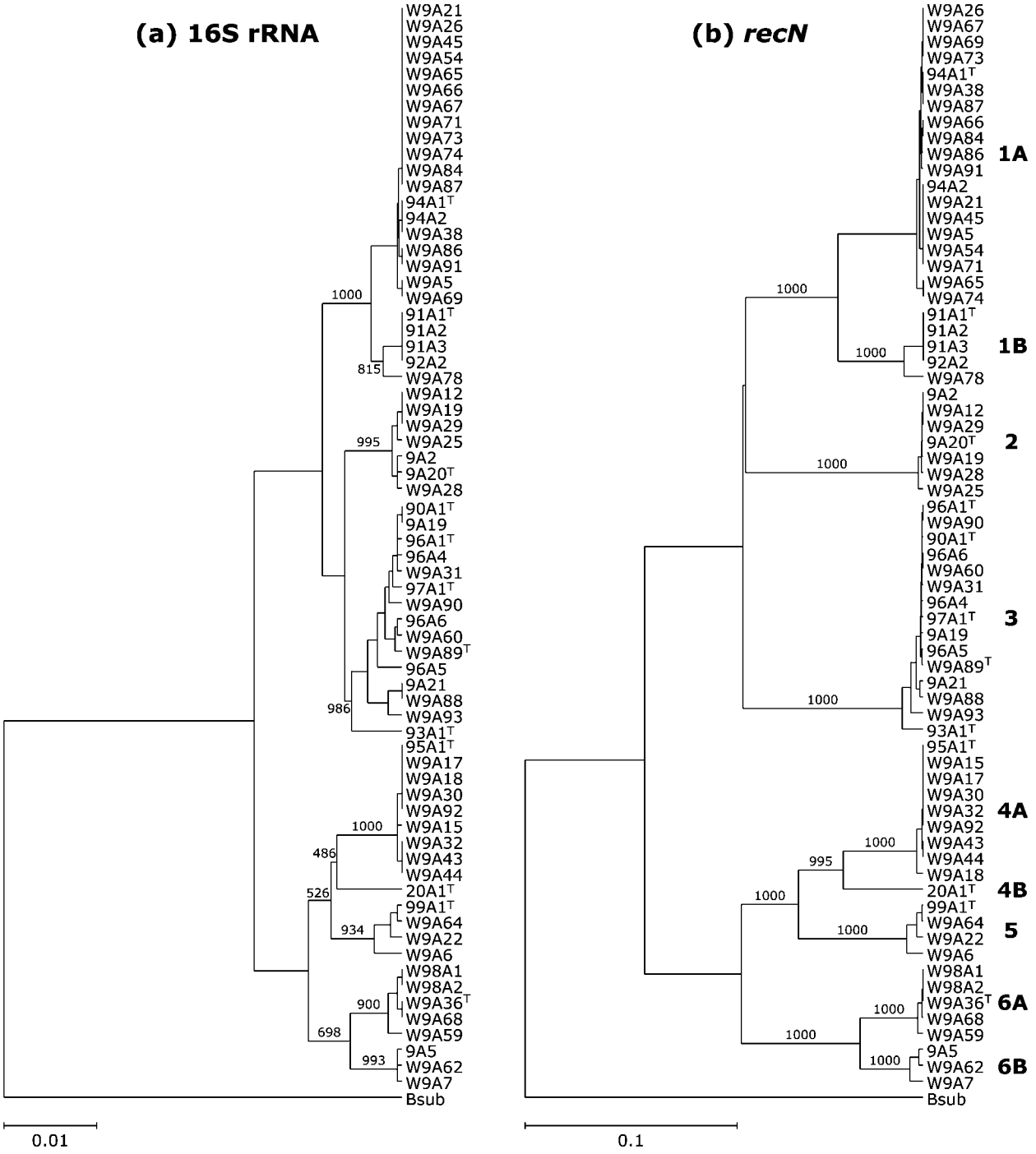

Fig. 1. Dendrograms showing the phylogenetic relationship among isolates of the genus *Geobacillus* based on full-length DNA sequences of (a) 16S rRNA genes and (b) *recN*. Bootstrap values (expressed as percentages of 1000 replications) are shown at major branching points. Possible taxonomic groupings suggested by *recN* analysis are indicated to the right of the figure. Strains are identified by the BGSC accession numbers listed in the Supplementary Table available in IJSEM Online. 'Bsub' refers to the genomically sequenced strain *Bacillus subtilis* 168 (GenBank/EMBL/DDBJ accession no. NC_000964). Bar, 1 substitution per (a) 100 nt and (b) 10 nt.

alignments versus the frequency of identical residues in 16S rRNA gene sequence alignments (Fig. 2). It is obvious from inspection that sequence identity scores for the two genes cluster tightly in the plot. The relationship between the identity scores fits a cubic equation with a high coefficient of determination ($R^2 = 95\,\%$), low standard error ($S = 2{\cdot}2\,\%$)

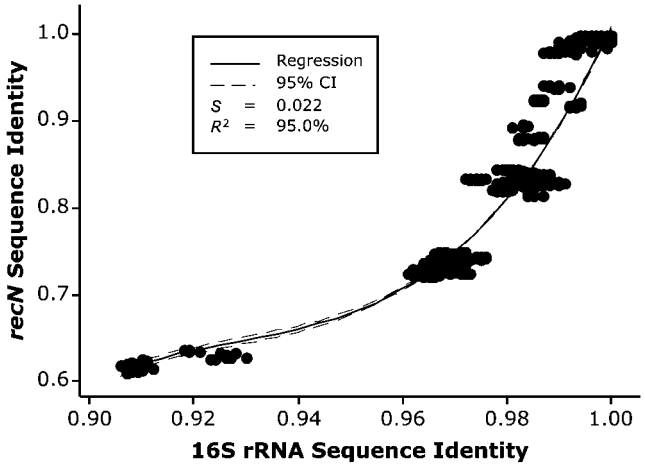

Fig. 2. Relationship between 16S rRNA gene and *recN* sequence identity scores for all pairwise combinations of strains listed in the Supplementary Table available in IJSEM Online. The solid line plots the cubic fit regression line, $y = -600{\cdot}8 + 1942x - 2091x^2 + 750{\cdot}9x^3$, where y is the *recN* sequence identity and x is the 16S rRNA sequence identity. The dashed line plots the 95 % confidence intervals for the regression line.

and narrow 95 % confidence intervals. The tight confidence intervals suggest that horizontal gene transfer has not disrupted the vertical co-transfer of the *recN* and 16S rRNA genes within this collection of *Geobacillus* strains. The exact parameters of this regression curve are unlikely to be of any special biological significance, but the higher-order equation (in this case, cubic) probably is significant. Within the genus *Geobacillus* – and probably among all bacteria that have retained the *recN* gene (Zeigler, 2003) – *recN* sequences have diverged much more rapidly than have 16S rRNA gene sequences. The linear portion of the plot in Fig. 2, where *recN* sequence identity is greater than about 85 % and 16S rRNA sequence identity is greater than about 98·5 %, has a slope of approximately 7·5. Among closely related *Geobacillus* strains, then, mutations are becoming fixed in *recN* at a rate almost an order of magnitude greater than they are in the 16S rRNA gene. It is well established, however, that as sequences continue to diverge, the variable positions become saturated with mutations (Gribaldo & Philippe, 2002). As a result, Fig. 2 reveals an inflection point where *recN* gene sequences become increasingly less reliable phylogenetic markers than 16S rRNA gene sequences. This comparison suggests that *recN* probably has a significantly greater resolving power than the 16S rRNA gene for assigning strains to taxa at the genus, species or subspecies level, but that at higher taxa *recN* might have considerably lower power.

To further test these concepts, 16S rRNA or *recN* gene sequences from five *Geobacillus* type strains were aligned with the corresponding sequences from 19 other members of the phylum *Firmicutes*, with the proteobacterium *Escherichia coli* K-12 as an outgroup. This group contains organisms that are related to one another at a wide variety of taxonomic levels. A comparison of the 16S rRNA gene and *recN* phylograms for the group, plotted at the same scale (Fig. 3), is instructive. For lower order taxa (genus, species and subspecies), *recN* has superior resolving power to the 16S rRNA gene, as reflected in the greater branch lengths separating the nodes that join organisms related at this level. In contrast, *recN* seems unsuitable for analysing higher order taxa (family, order and class); nodes joining organisms related at those levels are poorly separated with short branch lengths. This poor resolving power is probably related to mutational saturation in the rapidly diverging *recN* gene. For analysis of higher taxa, the more slowly diverging 16S rRNA gene appears to have a higher resolving power (Fig. 3).

The relationship between resolving power and mutational saturation at various phylogenetic depths can be quantified rather precisely with maximum-parsimony analysis, using the DNAPARS application of the PHYLIP phylogeny inference package (Felsenstein, 1989). From a set of aligned sequences, DNAPARS generates a tree based on the minimum number of mutational 'steps' required to account for the observed sequence differences for a proposed phylogeny. DNAPARS tallies the number of steps contributed by each individual residue position in the alignment. For a given alignment, a particular residue may be perfectly conserved, contributing zero steps and generating no phylogenetic signal. If a residue contributes one to five steps to the maximum-parsimony tree, it is generating a usable signal for constructing phylogenies. If it contributes a greater number of steps, it is generating noise, which becomes increasingly random as the position becomes saturated with mutations, confounding attempts to infer phylogenies (Gribaldo & Philippe, 2002). For this analysis, *recN* and 16S rRNA gene sequence alignments were produced for three strain sets of different phylogenetic depth. The '*Geobacillus thermoglucosidasius* set' contained all organisms falling in sequence identity groups 4A, 4B and 5 in Fig. 1, a level of similarity suggestive of closely related species or subspecies. The '*Geobacillus* set', containing a sample sequence from each of the nine similarity groups in Fig. 1, has a moderate phylogenetic depth, typical of a bacterial genus. The '*Firmicutes* set', containing the same species analysed in Fig. 2 (with *E. coli* omitted), has much more depth, typical of a bacterial phylum. Results of this analysis are presented in Table 1.

Table 1 confirms that *recN* is superior to the 16S rRNA gene at resolving lower taxa (genus and below), but that the 16S rRNA gene is much more useful for resolving higher taxa. With sets of closely related organisms, represented by the *G. thermoglucosidasius* and *Geobacillus* sets, neither gene generated significant noise. At the species–subspecies level, *recN* produced a phylogenetic signal roughly six times stronger than the 16S rRNA gene, with a correspondingly higher percentage of residues showing sequence variation. At the genus level, over 40 % of the *recN* residues generated a

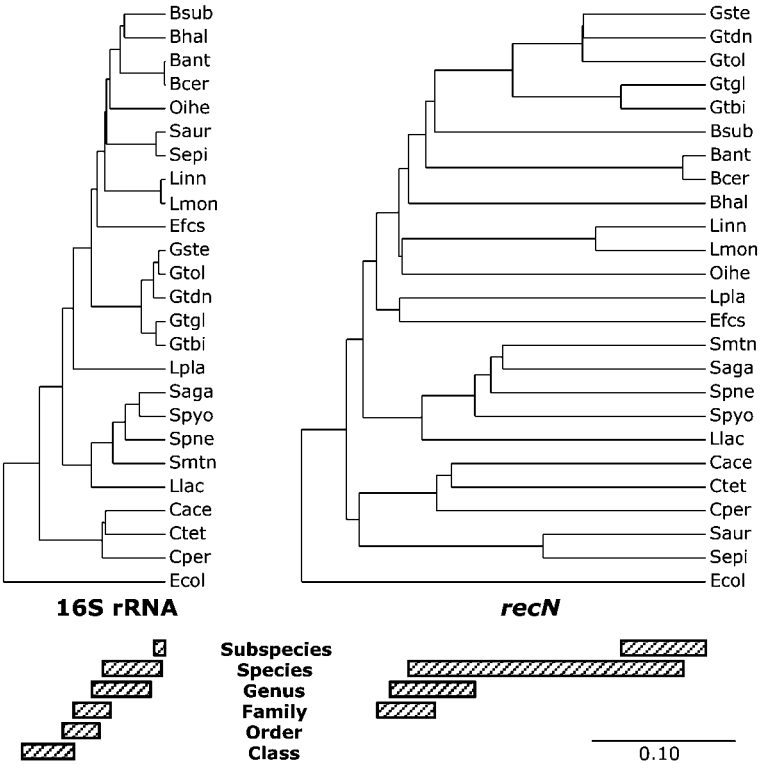

Fig. 3. Resolving power of 16S rRNA gene and *recN* phylogenies for elucidating bacterial relationships at various taxonomic ranks. Shaded rectangles indicate the approximate location of nodes joining bacteria of a given taxonomic rank in each phylogenetic tree. Bar, 1 substitution per 10 nt. Abbreviations: Bant, *Bacillus anthracis* Ames; Bcer, *B. cereus* ATCC 10987; Bhal, *B. halodurans* C125; Bsub, *B. subtilis* 168; Cace, *Clostridium acetobutylicum* ATCC 824^T; Cper, *C. perfringens* 13; Ctet, *C. tetani E88*; Ecol, *Escherichia coli* K-12; Efcs, *Enterococcus faecalis* V583; Gste, *G. stearothermophilus* BGSC 9A20^T; Gtbi, *G. toebii* BGSC 99A1^T; Gtdn, *G. thermodenitrificans* BGSC 94A1^T; Gtgl, *G. thermoglucosidasius* BGSC 95A1^T; Gtol, *G. thermoleovorans* BGSC 96A1^T; Linn, *Listeria innocua* Clip11262; Llac, *Lactococcus lactis* IL1403; Lmon, *Listeria monocytogenes* EGD-e; Lpla, *Lactobacillus plantarum* WCFS1; Oihe, *Oceanobacillus iheyensis* HTE831^T; Saga, *Streptococcus agalactiae* 2603VR; Saur, *Staphylococcus aureus* Mu50; Sepi, *Staphylococcus epidermidis* RP62A; Smtn, *Streptococcus mutans* UA159; Spne, *Streptococcus pneumoniae* R6; Spyo, *Streptococcus pyogenes* M1 GAS. *Geobacillus* sequences were determined in this work. All other sequences were taken from the publicly available GenBank genome sequences.

Table 1. Comparison of *recN* and 16S rRNA phylogenetic resolving power by maximum-parsimony analysis

The *G. thermoglucosidasius* set included each of the strains composing sequence identity groups 4A, 4B and 5 in Fig. 1. The *Geobacillus* set contained a sample strain from each of the sequence identity groups, including BGSC strains W9A19, 9A21, 9A5, W98A1, 91A1^T, 94A2, 20A1^T, 95A1^T and W9A6 (see the Supplementary Table in IJSEM Online). The *Firmicutes* set included the same group of species from the phylum *Firmicutes* analysed in Fig. 2. For a given strain set and gene alignment (16S rRNA gene or *recN*), the table gives the number of residues (with percentage of total residues in parentheses) that contribute no phylogenetic signal, good signal or noisy signal to a maximum-parsimony tree.

Set	Phylogenetic depth	Alignment	No. signal residues	Good signal		Noisy signal	
				Residues	Total steps	Residues	Total steps
G. thermoglucosidasius	Species–subspecies	16S rRNA	1518 (97 %)	45 (3 %)	53 (100 %)	0	0
		recN	1455 (84 %)	267 (16 %)	300 (100 %)	0	0
Geobacillus	Genus	16S rRNA	1467 (94 %)	98 (6 %)	148 (100 %)	0	0
		recN	988 (57 %)	734 (43 %)	1384 (100 %)	0	0
Firmicutes	Phylum	16 rRNA	900 (57 %)	608 (38 %)	1383 (70 %)	86 (5 %)	604 (30 %)
		recN	264 (15 %)	673 (38 %)	2125 (23 %)	831 (47 %)	6967 (77 %)

usable phylogenetic signal, as opposed to only 6 % of the 16S rRNA gene residues. Total signal strength was almost an order of magnitude greater for *recN* than for the 16S rRNA gene. At a phylum level, however, the 16S rRNA gene was clearly superior to *recN*. Although both genes began to show phylogenetically noisy residues at this level, 70 % of the signal produced by the 16S rRNA gene was still usable. In contrast, almost half of the *recN* residues were too highly saturated with mutations to generate usable signal, and 77 % of the total signal was of poor quality due to low signal-to-noise ratio. For the genus *Geobacillus* – and perhaps for other bacterial taxa as well – *recN* analysis should prove to be a powerful tool for organizing strains into lower taxa. The 16S rRNA gene does contribute enough useful phylogenetic signal even at this level, however, that alignments of concatenated *recN* and 16S rRNA gene sequences should further enhance the precision and accuracy of species and subspecies assignments.

Prediction of whole genome similarity by *recN* analysis

Zeigler (2003) suggested that *recN* sequence identity scores could predict, with a high degree of accuracy, the whole genome sequence identity shared by two organisms. From a survey of 44 complete genome sequences representing 16 bacterial genera, Zeigler (2003) developed the following model to relate SI_{genome}, the predicted DNA sequence identity shared by the genomes, and SI_{recN}, the sequence identity shared by their *recN* orthologues: $SI_{genome} = -1{\cdot}30 + 2{\cdot}25(SI_{recN})$.

The data from the current study could potentially allow an evaluation of this model for applicability to the genus *Geobacillus*. One useful comparison would be the predicted genome identity, as calculated from *recN* identity scores, with the percentage genome identity measured by DNA–DNA hybridization studies. Although the genus was only recently described, several references in the research literature do report DNA–DNA hybridization data for some of the strains in the present *Geobacillus* collection (Ahmad *et al.*, 2000; Caccamo *et al.*, 2000; Manachini *et al.*, 2000; Nazina *et al.*, 2001; Sung *et al.*, 2002; Sunna *et al.*, 1997; White *et al.*, 1993). Table 2 compares predicted with measured genome identity scores for these strains.

Table 2 shows that, in general, the Zeigler (2003) model predicts a somewhat higher genome identity than has been estimated from hybridization studies. The predicted and estimated values tend to be more similar towards their middle ranges and more dissimilar when comparing strains with very high or very low relatedness. It is difficult to assess whether these differences are due to inaccuracies in the *recN* prediction model, in the hybridization methodologies, or both. Table 2 does highlight a weakness of DNA–DNA hybridization studies; namely, the difficulty in reproducing similarity estimates obtained in different laboratories using different methods. Published estimates of genome similarity between *Geobacillus kaustophilus* and '*Bacillus caldotenax*', for example, range from 32 to 85 %, yielding conflicting answers to questions regarding the species identity of these bacteria. Nevertheless, it is entirely possible that the *recN* prediction model will require some recalibration for *Geobacillus*. In particular, the proposed species boundary of 89 % *recN* sequence identity (Zeigler, 2003) may be too conservative for this genus. At this point, it is not possible to use *recN* identity scores as the sole basis for assigning isolates to species within *Geobacillus*. It is clear, however, that *recN* comparisons could be a powerful tool

Table 2. Genome similarity values as predicted by *recN* sequence identity compared with those estimated by DNA hybridization methods

For each species comparison, percentage genome similarity was predicted from *recN* sequence alignments by the method of Zeigler (2003). Values in parentheses give the similarity estimates derived from published DNA hybridization studies. Strains or species: 1, *G. thermoleovorans*; 2, '*B. caldotenax*'; 3, '*Bacillus caldovelox*'; 4, *G. kaustophilus*; 5, *G. thermocatenulatus*; 6, *B. vulcani*; 7, *G. stearothermophilus*; 8, *G. subterraneus*; 9, *G. uzenensis*; 10, *G. thermodenitrificans*; 11, *G. thermoglucosidasius*; 12, *G. toebii*. References for hybridization data: Ahmad *et al.* (2000); Caccamo *et al.* (2000); Manachini *et al.* (2000); Nazina *et al.* (2001); Sung *et al.* (2002); Sunna *et al.* (1997); White *et al.* (1993).

Species	1	2	3	4	5	6	7	8	9	10	11
2	95 (82)										
3	95 (85)	94 (78)									
4	95 (84)	95 (32–85)	90 (75)								
5	91 (51–73)	90 (72)									
6		95 (51)		95 (61)							
7	56 (51)	57 (14)		57 (20–61)	56 (37)				59 (38)		
8	59 (48)				58 (50)		59 (53)		95 (49)		
9	59 (45)				58 (54)						
10	57 (21–31)			57 (40)	56 (47)	57 (41)	57 (32–48)	77 (5–12)	77 (45)		
11		34 (6)		34 (5)			35 (12–13)			37 (11–31)	
12											69 (27)

for the preliminary organization of isolates into possible taxa that could be validated by additional data from complementary methods.

Implications for *Geobacillus* taxonomy

During the brief period since its description, the genus *Geobacillus* (Nazina *et al.*, 2001) and its members have become a significant research focus. As Gram-positive thermophiles, these organisms have considerable potential for applications in biotechnology and bioremediation (Obojska *et al.*, 2002; Peng *et al.*, 2003). Their roles in natural and artificial thermal biotypes as well as in temperate soil environments are also of interest (Marchant *et al.*, 2002; McMullan *et al.*, 2004). The original genus description included eight species (Nazina *et al.*, 2001). The taxonomy of the group is in a rapid state of flux, however. On the one hand, the distinctiveness of several of these species has already been questioned (Sunna *et al.*, 1997). On the other hand, *Geobacillus* discovery programmes are uncovering novel isolates and spawning novel species proposals at a rapid rate (Banat *et al.*, 2004; Kuisiene *et al.*, 2004; Nazina *et al.*, 2004; Schäffer *et al.*, 2004). Analysis of *recN* gene sequences, in combination with 16S rRNA sequence analysis, could provide a powerful, high-throughput tool for validating and maintaining the taxonomy of this genus.

Both phylogenetic analyses represented in Fig. 1 cluster this set of thermophilic *Geobacillus* strains into nine similarity groups, all enjoying strong bootstrap support. Depending on where one chooses to draw the boundary demarcating inter- from intraspecific clusters, these homology groups could plausibly comprise from six to nine species. Groups 1A, 2, 4A, 5 and 6A appear to correspond unambiguously to the species *Geobacillus thermodenitrificans*, *G. stearothermophilus*, *G. thermoglucosidasius*, *Geobacillus toebii* and *Geobacillus caldoxylosilyticus*, respectively. Identification of the other four similarity groups with currently recognized species is somewhat more difficult, however.

Group 4B contains a single member, the proposed type strain of *Bacillus thermantarcticus* (Nicolaus *et al.*, 1996) (BGSC 20A1^T). The validity of this species has been questioned on technical grounds because, at the time of publication, the type strain was not deposited in two publicly accessible service collections in different countries (Euzéby & Tindall, 2004). An inspection of the original publication also suggests that the novel species was proposed on slender evidence. The authors reported no DNA–DNA hybridization data to test for genome similarity between *B. thermantarcticus* and related species. Their sole basis for distinguishing their novel isolate from the type strain of what is now termed *G. thermoglucosidasius*, which it closely resembled based on partial 16S rRNA gene sequence data, was a difference in G+C content (Nicolaus *et al.*, 1996). However, more recent measurements (Nazina *et al.*, 2001) show that the G+C content for *G. thermoglucosidasius*, as well as nearly every type strain in the genus *Geobacillus*, is virtually identical with those Nicolaus *et al.* (1996) reported for *B. thermantarcticus*. The *recN* and 16S rRNA gene sequence comparisons reported in this study form a sound basis for transferring this organism to the genus *Geobacillus*, either as a novel species or as a subspecies of *G. thermoglucosidasius*. Further analysis should readily distinguish between these possibilities.

Group 6B includes NUB3621 (=BGSC 9A5), doubtless the most well-characterized *Geobacillus* strain from a genetic standpoint. Systems for plasmid transformation (Wu & Welker, 1989), generalized transduction (Welker, 1988) and protoplast fusion (Chen *et al.*, 1986) have been described for this strain, and data generated from those studies have revealed a circular genetic map (Vallier & Welker, 1990). Although the research literature describes NUB3621 as *G. stearothermophilus*, the *recN* and 16S rRNA gene sequence analysis presented in Fig. 1 suggests that it is much more closely related to *G. caldoxylosilyticus*. It is probable that further analysis of Group 6B will result in the proposal of a novel subspecies of *G. caldoxylosilyticus* or of a new *Geobacillus* species.

The clustering of strains in similarity Group 3 raises significant questions for the taxonomy of the genus. Based on DNA–DNA hybridization data, Sunna *et al.* (1997) have suggested that the species described as *G. kaustophilus* and *G. thermocatenulatus* actually belong to *G. thermoleovorans*. In confirmation of their proposal, Group 3 includes the type strains of all three species (BGSC 90A1^T, BGSC 93A1^T and BGSC 96A1^T, respectively). Furthermore, the group also includes the type strains of *Bacillus vulcani* (Caccamo *et al.*, 2000) (BGSC 97A1^T) and of a recently proposed novel species, *Geobacillus lituanicus* (Kuisiene *et al.*, 2004) (BGSC W9A89^T). Clearly, a careful analysis of these species is required to clarify their relationships. It is interesting that all three of the *Geobacillus* strains currently the focus of genomic sequencing efforts – *G. stearothermophilus* strain 10 (equal to the BGSC 9A21 in Group 3), *G. kaustophilus* HTA426 and *G. thermoleovorans* T80 (McMullan *et al.*, 2004) – may be closely related. If the data analysed in Fig. 1 are representative of other members of these species, then one would predict that these three genome sequences will be found to differ only in detail.

Group 1B likewise presents a taxonomic puzzle. Its position on the *recN* and 16S rRNA gene phylograms (Fig. 1) suggests that this group either corresponds to a subspecies of *G. thermodenitrificans* (Group 1A) or composes a separate but closely related *Geobacillus* species. Indeed, the group contains the type strain of *Geobacillus subterraneus* (BGSC 91A1^T) along with two other isolates also described as belonging to that species (Nazina *et al.*, 2001). Yet the group also contains a strain described as *Geobacillus uzenensis* X (=BGSC 92A2) (Nazina *et al.*, 2001). The full-length 16S rRNA gene sequence determined for this strain in the present study (GenBank/EMBL/DDBJ accession no. AY608959) is only 98 % identical to the partial 16S rRNA gene sequence that served as the basis for its inclusion in *G. uzenensis* (GenBank/EMBL/DDBJ accession no. AF276305).

It is not clear whether sequencing errors in one or both GenBank entries account for the differences, or whether BGSC 92A2 is not in fact equivalent to strain X of Nazina *et al.* (2001). Although the type strain of *G. uzenensis* was not included in this study, its 16S rRNA gene sequence determined in our hands differs in only two to three positions from each of the sequences composing Group 3 in the present report (unpublished data). These data highlight the need for further analysis to confirm the taxonomic identity of *G. uzenensis* strains.

The study demonstrates the power of a highly variable but widely distributed sequence, such as *recN*, for organizing and maintaining the taxonomy of a bacterial genus. Further work should better calibrate *recN* as a molecular chronometer for *Geobacillus* and its relatives, allowing a more certain correlation between sequence identity scores and taxonomic relatedness. It appears that *recN* is a promising candidate for inclusion in a 'species prediction gene set' (Zeigler, 2003).

ACKNOWLEDGEMENTS

The author thanks the following individuals for their kind gift of bacterial strains: N. Bilgin, J. DiRuggiero, M. Hatsu, T. Kato, D. Mora, T. Nazina and C. Svenson. The author also thanks C. A. Dingus for assistance with statistical analysis, C. R. Zeigler for technical assistance, and the staff of the Plant–Microbe Genomics Facility at the Ohio State University for help with DNA sequencing. The Bacillus Genetic Stock Center (BGSC) is funded in part by a grant from the National Sciences Foundation (0234214).

REFERENCES

Ahmad, S., Scopes, R. K., Rees, G. N. & Patel, B. K. C. (2000). *Saccharococcus caldoxylosilyticus* sp. nov., an obligately thermophilic, xylose-utilizing, endospore-forming bacterium. *Int J Syst Evol Microbiol* **50**, 517–523.

Banat, M., Marchant, R. & Rahman, T. J. (2004). *Geobacillus debilis* sp. nov., a novel obligately thermophilic bacterium isolated from a cool soil environment, and reassignment of *Bacillus pallidus* to *Geobacillus pallidus* comb. nov. *Int J Syst Evol Microbiol* **54**, 2197–2201.

Caccamo, D., Gugliandolo, C., Stackebrandt, E. & Maugeri, T. L. (2000). *Bacillus vulcani* sp. nov., a novel thermophilic species isolated from a shallow marine hydrothermal vent. *Int J Syst Evol Microbiol* **50**, 2009–2012.

Chen, Z. F., Wojcik, S. F. & Welker, N. E. (1986). Genetic analysis of *Bacillus stearothermophilus* by protoplast fusion. *J Bacteriol* **165**, 994–1001.

Coenye, T., Falsen, E., Vancanneyt, M., Hoste, B., Govan, J. R. W., Kersters, K. & Vandamme, P. (1999). Classification of *Alcaligenes faecalis*-like isolates from the environment and human clinical samples as *Ralstonia gilardii* sp. nov. *Int J Syst Bacteriol* **49**, 405–413.

Edwards, U., Rogall, T., Blocker, H., Emde, M. & Bottger, E. C. (1989). Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res* **17**, 7843–7853.

Euzéby, J. P. & Tindall, B. J. (2004). Status of strains that contravene Rules 27(3) and 30 of the Bacteriological Code. Request for an Opinion. *Int J Syst Evol Microbiol* **54**, 293–301.

Felsenstein, J. (1989). PHYLIP – Phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166.

Gribaldo, S. & Philippe, H. (2002). Ancient phylogenetic relationships. *Theor Popul Biol* **61**, 391–408.

Gürtler, V. & Mayall, B. C. (2001). Genomic approaches to typing, taxonomy and evolution of bacterial isolates. *Int J Syst Evol Microbiol* **51**, 3–16.

Kuisiene, N., Raugalas, J. & Chitavichius, D. (2004). *Geobacillus lituanicus* sp. nov. *Int J Syst Evol Microbiol* **54**, 1991–1995.

Lu, Z., Liu, Z., Wang, L., Zhang, Y., Qi, W. & Goodfellow, M. (2001). *Saccharopolyspora flava* sp. nov. and *Saccharopolyspora thermophila* sp. nov., novel actinomycetes from soil. *Int J Syst Evol Microbiol* **51**, 319–325.

Manachini, P. L., Mora, D., Nicastro, G., Parini, C., Stackebrandt, E., Pukall, R. & Fortina, M. G. (2000). *Bacillus thermodenitrificans* sp. nov., nom. rev. *Int J Syst Evol Microbiol* **50**, 1331–1337.

Marchant, R., Banat, I. M., Rahman, T. J. & Berzano, M. (2002). The frequency and characteristics of highly thermophilic bacteria in cool soil environments. *Environ Microbiol* **4**, 595–602.

McMullan, G., Christie, J. M., Rahman, T. J., Banat, I. M., Ternan, N. G. & Marchant, R. (2004). Habitat, applications and genomics of the aerobic, thermophilic genus *Geobacillus*. *Biochem Soc Trans* **32**, 214–217.

Nazina, T. N., Tourova, T. P., Poltaraus, A. B. & 8 other authors (2001). Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillus subterraneus* gen. nov., sp. nov. and *Geobacillus uzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus* as the new combinations *G. stearothermophilus*, *G. thermocatenulatus*, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoglucosidasius* and *G. thermodenitrificans*. *Int J Syst Evol Microbiol* **51**, 433–446.

Nazina, T. N., Lebedeva, E. V., Poltaraus, A. B., Tourova, T. P., Grigoryan, A. A., Sokolova, D. Sh., Lysenko, A. M. & Osipov, G. A. (2004). *Geobacillus gargensis* sp. nov., a novel thermophile from a hot spring, and the reclassification of *Bacillus vulcani* as *Geobacillus vulcani* comb. nov. *Int J Syst Evol Microbiol* **54**, 2019–2024.

Nicolaus, B., Lama, L., Esposito, E., Cristina Manca, M. C., di Prisco, G. & Gambacorta, A. (1996). "*Bacillus thermoantarcticus*" sp. nov., from Mount Melbourne, Antarctica: a novel thermophilic species. *Polar Biol* **16**, 101–104.

Obojska, A., Ternan, N. G., Lejczak, B., Kafarski, P. & McMullan, G. (2002). Organophosphonate utilization by the thermophile *Geobacillus caldoxylosilyticus* T20. *Appl Environ Microbiol* **68**, 2081–2084.

Peng, X., Misawa, N. & Harayama, S. (2003). Isolation and characterization of thermophilic bacilli degrading cinnamic, 4-coumaric, and ferulic acids. *Appl Environ Microbiol* **69**, 1417–1427.

Schäffer, C., Franck, W. L., Scheberl, A., Kosma, P., McDermott, T. R. & Messner, P. (2004). Classification of isolates from locations in Austria and Yellowstone National Park as *Geobacillus tepidamans* sp. nov. *Int J Syst Evol Microbiol* **54**, 2361–2368.

Stackebrandt, E. & Goebel, B. M. (1994). Taxonomic note: 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., Frederiksen, W., Garrity, G. M. & 10 other authors (2002). Report of the ad hoc committee for the re-evaluation of

the species definition in bacteriology. *Int J Syst Evol Microbiol* **52**, 1043–1047.

Sung, M.-H., Kim, H., Bae, J.-W. & 9 other authors (2002). *Geobacillus toebii* sp. nov., a novel thermophilic bacterium isolated from hay compost. *Int J Syst Evol Microbiol* **52**, 2251–2255.

Sunna, A., Tokajian, S., Burghardt, J., Rainey, F., Antranikian, G. & Hashwa, F. (1997). Identification of *Bacillus kaustophilus*, *Bacillus thermocatenulatus* and *Bacillus* strain HSR as members of *Bacillus thermoleovorans*. *Syst Appl Microbiol* **20**, 232–237.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Vallier, H. & Welker, N. E. (1990). Genetic map of the *Bacillus stearothermophilus* NUB36 chromosome. *J Bacteriol* **172**, 793–801.

Wayne, L. G., Brenner, D. J., Colwell, R. R. & 9 other authors (1987). International Committee on Systematic Bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Bacteriol* **37**, 463–464.

Welker, N. E. (1988). Transduction in *Bacillus stearothermophilus*. *J Bacteriol* **170**, 3761–3764.

White, D., Sharp, R. J. & Priest, F. G. (1993). A polyphasic taxonomic study of thermophilic bacilli from a wide geographical area. *Antonie van Leeuwenhoek* **64**, 357–386.

Wu, L. J. & Welker, N. E. (1989). Protoplast transformation of *Bacillus stearothermophilus* NUB36 by plasmid DNA. *J Gen Microbiol* **135**, 1315–1324.

Zeigler, D. R. (2003). Gene sequences useful for predicting relatedness of whole genomes in bacteria. *Int J Syst Evol Microbiol* **53**, 1893–1900.