APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Mar. 2005, p. 1328–1335 Vol. 71, No. 3
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.3.1328–1335.2005

Rhizobial 16S rRNA and *dnaK* Genes: Mosaicism and the Uncertain Phylogenetic Placement of *Rhizobium galegae*

B. D. Eardly,[1]* S. M. Nour,[2] P. van Berkum,[3] and R. K. Selander[4]

Pennsylvania State University, Berks Campus, Reading,[1] and Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park,[4] Pennsylvania; Soybean Genomics and Improvement Laboratory, USDA Agricultural Research Service, Beltsville, Maryland[3]; and Agriculture and Agri-food Canada, London, Ontario, Canada[2]

Received 22 June 2004/Accepted 13 October 2004

The phylogenetic relatedness among 12 agriculturally important species in the order *Rhizobiales* was estimated by comparative 16S rRNA and *dnaK* sequence analyses. Two groups of related species were identified by neighbor-joining and maximum-parsimony analysis. One group consisted of *Mesorhizobium loti* and *Mesorhizobium ciceri*, and the other group consisted of *Agrobacterium rhizogenes*, *Rhizobium tropici*, *Rhizobium etli*, and *Rhizobium leguminosarum*. Although bootstrap support for the placement of the remaining six species varied, *A. tumefaciens*, *Agrobacterium rubi*, and *Agrobacterium vitis* were consistently associated in the same subcluster. The three other species included *Rhizobium galegae*, *Sinorhizobium meliloti*, and *Brucella ovis*. Among these, the placement of *R. galegae* was the least consistent, in that it was placed flanking the *A. rhizogenes-Rhizobium* cluster in the *dnaK* nucleotide sequence trees, while it was placed with the other three *Agrobacterium* species in the 16S rRNA and the DnaK amino acid trees. In an effort to explain the inconsistent placement of *R. galegae*, we examined polymorphic site distribution patterns among the various species. Localized runs of nucleotide sequence similarity were evident between *R. galegae* and certain other species, suggesting that the *R. galegae* genes are chimeric. These results provide a tenable explanation for the weak statistical support often associated with the phylogenetic placement of *R. galegae*, and they also illustrate a potential pitfall in the use of partial sequences for species identification.

As with most bacteria, evolutionary relationships among the members of the order *Rhizobiales* are usually estimated through 16S rRNA sequence comparisons. One problem with this approach however, is that these estimates are not always congruent with estimates derived from other loci (8, 23). In this study we obtained 16S rRNA and *dnaK* sequences for 12 agriculturally important species that were chosen to represent three families (*Rhizobiaceae*, *Phyllobacteriaceae*, and *Brucellaceae*) within the order *Rhizobiales*. Our goal was to compare estimates of phylogenetic relatedness based on their 16S rRNA sequences with corresponding estimates based on their *dnaK* sequences.

Within the family *Rhizobiaceae*, there are two closely related genera of agricultural significance: the nitrogen-fixing mutualists of the genus *Rhizobium* and the plant pathogens of the genus *Agrobacterium*. Young et al. (30) recently proposed that the genus *Rhizobium* be emended to include the species within the genus *Agrobacterium*. Part of the justification for this proposal was the observation that the 16S rRNA sequences of certain species of *Rhizobium* (e.g., *R. galegae)* are more similar to *Agrobacterium* sequences than they are to *Rhizobium* sequences. This proposal has been controversial, however, because of the ecological and genomic differences that exist between the two genera (6, 31).

Among the 12 species we examined, 4 were nitrogen-fixing species of the genus *Rhizobium* (*R. galegae*, *R. leguminosarum*, *R. tropici*, and *R. etli*) and 4 were phytopathogenic species of the genus *Agrobacterium* (*A. tumefaciens*, *A. rhizogenes*, *A. rubi*, and *A. vitis*). The remainder were representatives of three more distantly related genera, including two nitrogen-fixing species of the genus *Mesorhizobium*, a nitrogen-fixing species of the genus *Sinorhizobium*, and one mammalian pathogen of the genus *Brucella*. Since published *dnaK* sequences (3, 5, 18) were only available for 3 of the 12 species, it was necessary to clone and sequence *dnaK* genes from the other 9.

The *dnaK* gene encodes a highly conserved chaperone protein that performs multiple functions in the cell, including the folding of nascent polypeptides, the assembly and disassembly of multimeric protein structures, membrane translocation of secreted proteins, and the degradation of proteins (1, 9, 10). Furthermore, DnaK, the prokaryotic homolog of the eukaryotic Hsp70 protein, has been found in all eubacterial species examined to date. The DnaK proteins of *Bacillus subtilis* and *Escherichia coli* have three functionally distinct domains, including an N-terminal ATPase-binding domain, a substrate-binding domain, and a small C-terminal domain that has been associated with the degradation of the σ^{32} subunit of RNA polymerase (13). Both the ATPase and substrate-binding domains of *dnaK* are highly conserved and appear to form a species-specific functional unit, while the C-terminal domain of *dnaK* is less well conserved. Apparently the C-terminal domain can maintain functional activity when exchanged between *E. coli* and *B. subtilis* (13).

In this report, we describe the results of a phylogenetic analysis of the 16S and *dnaK* genes of *Brucella ovis*, *Mesorhizobium ciceri*, *Mesorhizobium loti*, *R. etli*, *R. galegae*, *R. leguminosarum*, *A. tumefaciens*, *A. rhizogenes*, *A. rubi*, *R. tropici*, *A. vitis*, and *Sinorhizobium meliloti*. Because the analysis revealed a lack of congruence in the placement of certain species in the

* Corresponding author. Mailing address: Pennsylvania State University, Berks Campus, P.O. Box 7009, Reading, PA 19610. Phone: (610) 396-6131. Fax: (610) 396-6024. E-mail: bde1@psu.edu.

TABLE 1. Sequences of *dnaK* primers used in this study

Primer	Sequence(5′→3′)
Forward	
KF1	ATYGGNATYGAYCTNGGNAC
KF1modB	ATYGGTATYGACCTKGGMAC
HSF1mod	GACCTGGGCACGACCAACTC
dnak5F	GACCGAAATCAACCTGCC
dnak6F	GACTGAAATCAACCTGCC
dnak1F	GGTGAAGACTTCGACAT
dnak23F	GAAGATCGAACTGTCCTC
dnak23.1F	CAAGATCGAGCTGTCGTC
STR-1	GCTCCTTATATACGCCGCA
STR-2	GGTTCCGACARCTYGCTT
STR-3	CTYGCTTCAAGGAGAGA
Reverse	
KR2	TCRAANGTNACYTCDATYTG
KR2modB	TCGAASGTSACYTCGATCTG
4.3mod	ATTGGCTTCGGCGTCCTT
dnaJ1R	TCGTAGGCYTCGTTGAT
dnaJ4R	AAGGCGCTYTTCAGCTCT
dnaJ9R	TCGTCCTTGATTTCCTCG
HL-2R	TGCTTCACGATCTCCTGG

16S rRNA and *dnaK* trees, we also examined the sequences for evidence of intragenic recombination.

MATERIALS AND METHODS

Bacterial strains. Nucleotide sequences for *dnaK* were obtained from the following nine rhizobial type strains: *A. rhizogenes* ATCC 11325^T, *A. rubi* ATCC 13335^T, *A. vitis* ATCC 49767^T, *R. etli* ATCC 51251^T, *R. galegae* ATCC 43677^T, *R. leguminosarum* ATCC 10004^T, *R. tropici* ATCC 49672^T, *M. ciceri* ATCC 51585^T, and *M. loti* ATCC 33669^T. All strains were grown and maintained in yeast extract-mannitol medium (25).

DNA extraction and PCR amplification. DNA extractions were performed by standard methods (16). The sequences of the oligonucleotide primers used for PCR amplification are listed in Table 1. These corresponded to conserved regions observed in the published *dnaK* sequences for *S. meliloti* and *A. tumefaciens*. Initially, only 1.4 kb of the 1.9-kb *dnaK* genes was cloned and sequenced. Building upon these core segments, gene-walking primers were then designed to amplify (and clone) upstream (100 bp) and downstream (400 bp) *dnaK* segments (Table 2). All primers were synthesized with a Beckman 1000 oligonucleotide synthesizer (Beckman, Fullerton, Calif.). Template DNA for cloning and sequencing was obtained through a 30-cycle amplification series as follows: 94°C for 1 min, 55°C for 45 s, and 72°C for 2 min, with an initial denaturing step of 94°C for 5 min. The annealing temperature was adjusted for increased stringency. Bands of interest were excised from low-melting-point agarose gels (1.5%) and purified with the protocol from the QIAGEN gel extraction kit (QIAGEN, Inc., Valencia, Calif.).

Cloning and nucleotide sequencing. Purified PCR products were cloned into the pPCR-ScriptTM Amp SK(+) cloning vector, and transformation was done according to the manufacturer's instructions (Stratagene, La Jolla, Calif.). Transformants were examined for the presence of a recombinant plasmid containing the desired inserts by using the ScreenTest recombinant screening kit (Stratagene). After verification, the selected recombinant plasmids were extracted with the QIAprep miniprep kit (QIAGEN, Inc.), and sequences were obtained with standard sequencing primers supplied by the manufacturer. Both the positive and negative strands of the inserted DNA were sequenced three times to eliminate possible sequencing errors. Sequences were obtained by using a dye-deoxy cycle sequencing kit in combination with an ABI Prism model 373 automated sequencer (Perkin-Elmer Applied Biosystems, Foster City, Calif.). Electropherograms were compared for accuracy of the results, and complete sequences were obtained by assembling overlapping contigs with DNASTAR (DNASTAR, Inc., Madison, Wis.).

Phylogenetic analysis. The 16S rRNA sequences obtained from the database corresponded to positions 29 through 1491 in the *rrnB* sequence of *E. coli* (2). Alignments of the sequences were obtained with the ClustalW (21) program and by inspection. Parameters for the ClustalW program included the slow-accurate alignment parameter, the IUB DNA weight matrix, and (for protein sequences) the PAM 250 protein weight matrix. The *dnaK* nucleotide sequences corresponded to positions 442 through 2352 in the published sequence of *B. ovis* (3). These were aligned by a two-step process. In the first step, ClustalW and inspection were used to align the inferred DnaK amino acid sequences. In the second step, the DnaK amino acid sequence alignment was used as input for the CodonAlign 2.0 nucleotide sequence alignment program (http://sinauer.com/hall/). This program generates a nucleotide sequence alignment containing triplet gaps that correspond to the codon gaps in the amino acid alignment. Neighbor-joining phylogenies for the aligned data sets were constructed by using Jukes-Cantor distances. Maximum-parsimony trees for the aligned data sets were generated with a heuristic min-mini tree search option with a search factor of 2. Bootstrap confidence levels were based on 1,000 permutations of the data sets. Software implementations of these programs were available in MEGA, version 2.1 (11). Discordance between 16S rRNA and *dnaK* phylogenies was tested with the incongruence length difference (ILD) test (7). In this test, the number of steps necessary for minimum-length trees in separate (partitioned) analyses is calculated and then the incongruence between separate data matrices is measured by the additional steps required when the separate matrices are combined into a single analysis. If the summed length of the combined trees is significantly longer than that of the original trees ($P < 0.05$), more incongruence is present between the two sets of data than can be explained by chance alone.

Statistical tests for recombination. Two contrasting nucleotide substitution-based methods were used to evaluate the distribution of polymorphic nucleotides in the sequences (17, 20). The Stephens test (20) classifies each polymorphic position (or column) in a sequence alignment according to how the nucleotides at that position partition the sequences into phylogenetic groups. The software for this analysis is available at http://www.shigatox.net/stec/index.html. The Stephens test generates statistics for each possible phylogenic partition, including the following: d_o, which is the observed number of nucleotides between the most widely spaced sites that support a particular partition, and g_o, which is the number of nucleotides between consecutive sites that support a particular partition. The program calculates the probability that the distance (d) between a random pair of sites is less than or equal to the observed distance (d_o) between two sites supporting a particular partition [P ($d \leq d_o$)]. Improbably small d_o values were taken as evidence of clustering. In contrast, g_o values that were improbably large [e.g., P ($g \geq g_o$) = 0.01] were taken as evidence of significant gaps between sites supporting a particular partition. In our analysis, partitions that were supported by less than five consecutive polymorphic positions in the alignment were considered to be trivial and were excluded from further analyses. Because the levels of statistical significance associated with recombinant segments could have been influenced by adjacent hypervariable (or nonvariable) regions (20), significant Stephens test statistics were subsequently validated by the stepwise removal of adjacent hypervariable (or nonvariable) regions followed by retesting. The purpose of this additional procedure was to confirm that

TABLE 2. *dnaK* sequencing primers used in this study

Strain	1,400-bp sequence	5′ end	3′ end
M. loti	HSF1mod/4.3mod	STR-3/HL-2R	dnak5F/dnaJ1R
R. galegae	KF1modB/KR2modB	STR-1/KR2modB	STR-1/dnaJ9R
A. vitis	KF1/KR2	STR-3/KR2modB	dnak6F/dnaJ1R
A. rhizogenes	KF1modB/KR2modB	STR-2/KR2modB	dnak23F/dnaJ9R
A. rubi	KF1/KR2	STR1/KR2modB	dnak5F/dnaJ4R
R. etli	KF1modB/KR2modB	STR1/KR2modB	dnak5F/dnaJ1R
R. tropici	KF1/KR2	STR3/KR2modB	dnak1F/dnaJ1R
R. leguminosarum	KF1modB/KR2modB	STR1/KR2modB	dnak1F/dnaJ1R
M. ciceri	HSF1mod/4.3mod	STR3/HL-2R	dnak23.1F/dnaJ1R

recombinant segments had not been identified solely on the basis of their proximity to hypervariable (or nonvariable) regions.

Sawyer's Geneconv method (17) also was used to analyze the distribution of polymorphic positions in the sequence alignments. The software for this program was obtained from http://www.math.wustl.edu/~sawyer/geneconv. This method is used to identify statistically significant runs of sequence similarity through multiple pairwise sequence comparisons. Empirical significance levels for this method are obtained by comparing statistics associated with runs of polymorphic sites with corresponding statistics for 10,000 random permutations of the sites. Unlike the Stephens test, the sum-of-square scores in Sawyer's method are unaffected by mutational hot (or cold) spots (17). The global permutation *P* values reported are the proportion of run scores in the simulation whose lengths exceeded the observed values for the random permutations. Because segments involved in recombination may be affected by later mutation, the stringent requirements for absolute sequence identity were relaxed in secondary assessments through the implementation of a mismatch scoring penalty.

Nucleotide sequence accession numbers. The GenBank accession numbers for the *dnaK* nucleotide sequences included in this study are as follows: *A. rhizogenes*, AY752737; *A. rubi*, AY752738; *A. vitis*, AY752739; *A. tumefaciens*, X87113; *B. ovis*, M94063; *M. ciceri*, AY752740; *M. loti*, AY752741; *R. etli*, AY752742; *R. galegae*, AY752743; *R. leguminosarum*, AY752744; *R. tropici*, AY752745; and *S. meliloti*, L36602. The sequences with an "AY" prefix (above) have not been reported previously. The GenBank accession numbers of the 16S rRNA sequences obtained from the database are as follows: *A. rhizogenes*, X67224; *A. rubi*, X67228; *A. vitis*, X67225; *A. tumefaciens*, X67223; *B. ovis*, L26168; *M. ciceri*, U07934; *M. loti*, X67229; *R. etli*, U28916; *R. galegae*, X67226; *R. leguminosarum*, X67227; *R. tropici*, X67233; and *S. meliloti*, X67222.

RESULTS

DNA sequences. The nine *dnaK* sequences determined in this study were aligned with the three published rhizobial *dnaK* sequences. This alignment resulted in 1,902 bp of *dnaK* sequence for the analysis. Corresponding 16S rRNA sequences for the same 12 species resulted in a 1,401-bp alignment for analysis.

Phylogenetic tree comparisons and analysis. Trees generated by neighbor-joining and maximum-parsimony methods had similar topologies (Fig. 1 and 2). *A. rhizogenes*, *R. tropici*, *R. etli*, and *R. leguminosarum* were all placed in a single group (the *Rhizobium* clade) that was supported by high bootstrap percentages. With the exception of the two *Mesorhizobium* species, the placement of the other species varied with respect to each other, as well as with respect to their position relative to the *Rhizobium* clade. The lack of congruence between the 16S rRNA and *dnaK* nucleotide sequence trees was evaluated by the ILD test. The incongruence was significant ($P = 0.01$), indicating that the two nucleotide data sets (16S rRNA and *dnaK*) provided different phylogenetic signals. Because the ILD test does not accommodate a combination of both nucleotide and amino acid sequence data, the DnaK amino acid sequence data were not included in the ILD analysis.

In both of the *dnaK* nucleotide sequence trees, *R. galegae* was placed flanking the *Rhizobium* clade (Fig. 1B and 2B). In contrast, in both the 16S rRNA and DnaK amino acid trees (Fig. 1A and 2A and 1C and 2C, respectively), *R. galegae* was placed with the members of the *Agrobacterium* clade (*A. tumefaciens*, *A. rubi*, and *A. vitis*). In an effort to explain the inconsistent placement of this particular species, as well as the generally low level of statistical support associated with its placement (with the exception of Fig. 2A), the sequences of *R. galegae* were compared to those of members of the *Agrobacterium* and *Rhizobium* clades.

The percent sequence identity between the 16S rRNA genes of *R. galegae* and those of *A. rubi*, *A. tumefaciens*, and *A. vitis* averaged 95.6%, with values ranging from 95.2 to 96.1%. Cor-

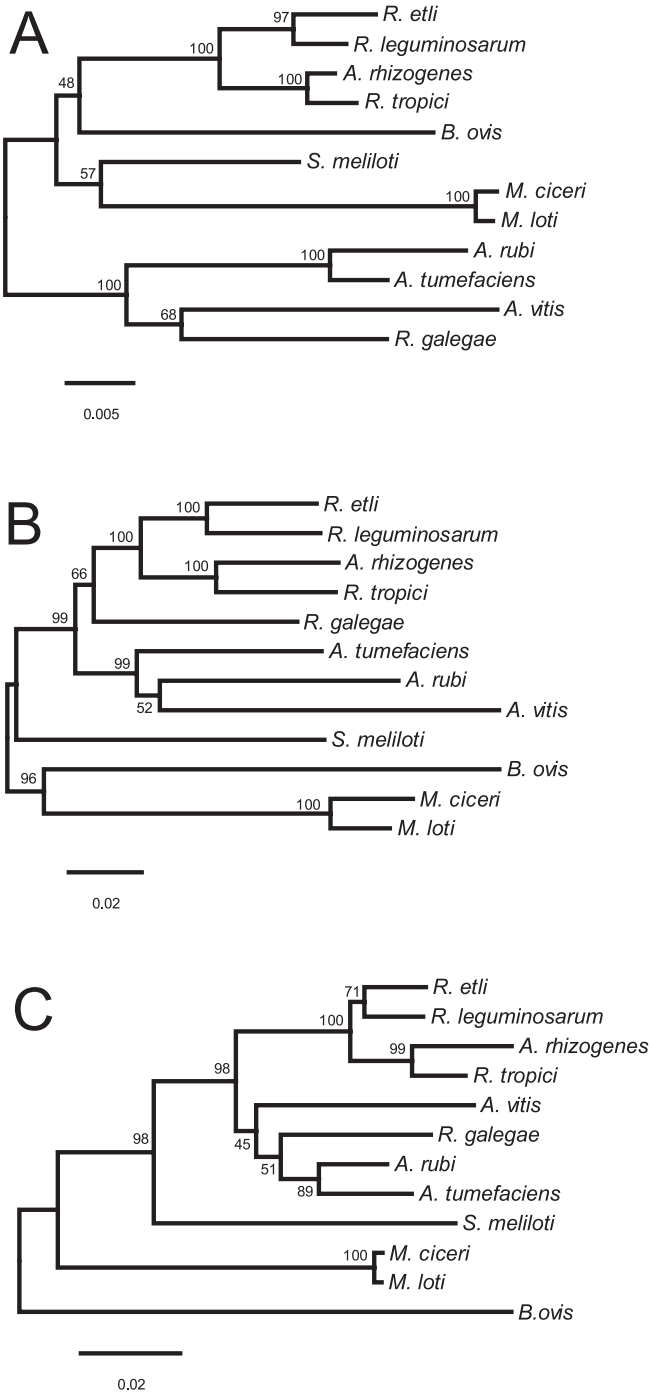

FIG. 1. Unrooted neighbor-joining trees based on single-gene sequences from 12 species representing five genera (*Agrobacterium*, *Brucella*, *Mesorhizobium*, *Rhizobium*, and *Sinorhizobium*) within the order *Rhizobiales*. Percentage bootstrap support at each internal node is based on 1,000 replicate trees. (A) 16S rRNA nucleotide sequence tree. The total alignment length for the analysis was 1,401 bp. Positions with gaps were omitted, and the Kimura two-parameter distance correction was applied. (B) *dnaK* nucleotide sequence tree. The total alignment length for the analysis was 1,917 bp. Positions with gaps were omitted, and the Kimura two-parameter distance correction was applied. (C) Inferred DnaK amino acid sequence tree. The total alignment included 642 amino acid residues. Positions with gaps were omitted, and rates of amino acid substitution were assumed to follow a Poisson distribution.

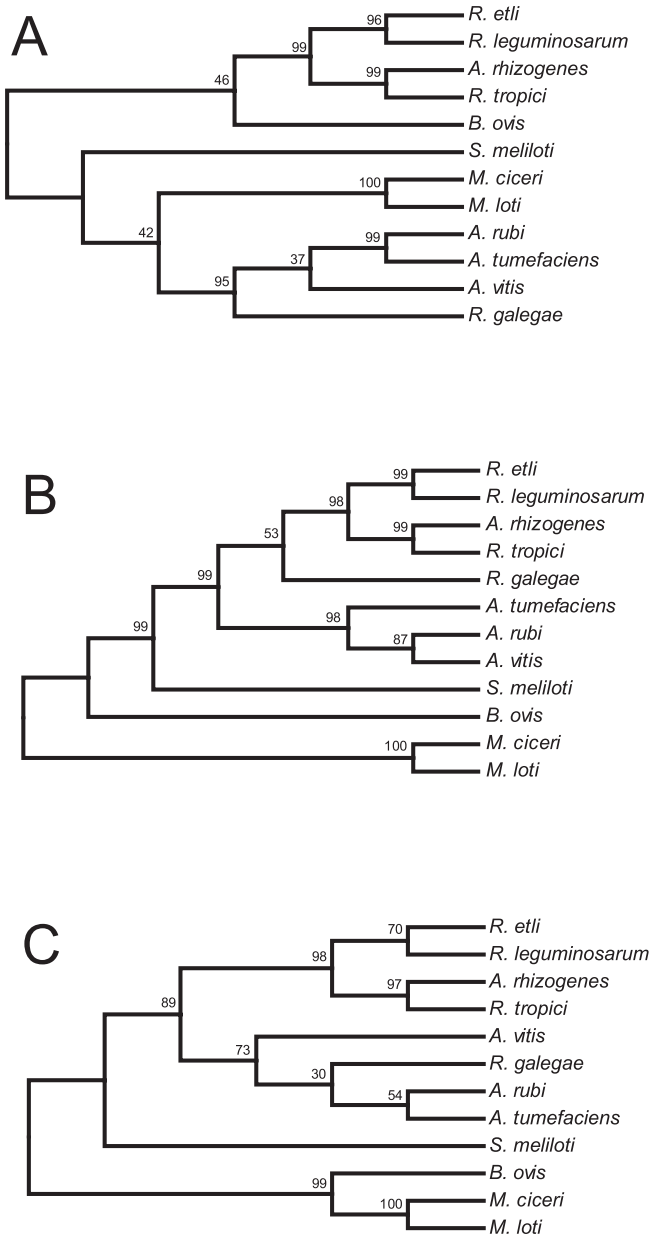

FIG. 2. Unrooted maximum-parsimony trees based on single-gene sequences from 12 species representing five genera (*Agrobacterium*, *Brucella*, *Mesorhizobium*, *Rhizobium*, and *Sinorhizobium*) within the order *Rhizobiales*. Trees were generated by a heuristic mini-min tree search option with a search factor of 2. The percentage bootstrap support at each internal node is based on 1,000 replicate trees. (A) 16S rRNA nucleotide sequence tree. The total alignment length for the analysis was 1,401 bp. (B) *dnaK* nucleotide sequence tree. The total alignment length for the analysis was 1,917 bp. (C) Inferred DnaK amino acid sequence tree. The total alignment included 642 amino acid residues.

respondingly, the percent sequence identity between the 16S rRNA genes of *R. galegae* and those of *R. etli*, *R. leguminosarum*, *R. tropici*, and *A. rhizogenes* was slightly less (95.2%), with values ranging from 94.6 to 95.9%. Particularly noteworthy, however, was the relatively high percentage of sequence identity between the *R. galegae* 16S rRNA sequence and that of *R. leguminosarum* (95.9%). Only *A. vitis* shared a higher level of sequence identity with *R. galegae* (96.1%).

TABLE 3. Number of nucleotide substitution differences between the *dnaK* genes of *R. galegae* and representatives of the *Agrobacterium* and *Rhizobium* clades

Sequence for comparison	No. of differences			
	Entire *dnaK* gene		*dnaK* segment 862–1178[a]	
	Synonymous	Nonsynonymous	Synonymous	Nonsynonymous
A. rubi	204	63	37	3
R. tropici	130	67	13	1

[a] Segment beginning at nucleotide position 862 and ending at position 1178 and corresponding to amino acid positions 288 through 393 in Fig. 3.

A similar pattern emerged in parallel comparisons among the DnaK amino acid sequences. The average percent sequence identity between the *R. galegae* DnaK amino acid sequence and those of the *Agrobacterium* clade was 93.1%, with values ranging from 91.0 to 94.1%. The average percent sequence identity of the *R. galegae* DnaK sequence to those of the *Rhizobium* clade was 91.9%, slightly less than the average for the *Agrobacterium* clade sequences. As in the 16S rRNA comparisons, the *R. galegae* DnaK amino acid sequence had a relatively high level of percent sequence identity to the DnaK amino acid sequence of *R. leguminosarum* (93.1%).

Even though the nucleotide sequence of the *dnaK* gene in *R. galegae* was more similar to those of members of the *Rhizobium* clade, averaging 88.2% sequence identity, it shared a relatively high level of identity with the sequence of *A. tumefaciens* (88.3%). In contrast, the *R. galegae dnaK* nucleotide sequence of *R. galegae* shared only 85.6 and 84.2% sequence identities, respectively, with the *dnaK* nucleotide sequences of *A. rubi* and *A. vitis*.

In an effort to further explain the placement of *R. galegae* flanking the *Rhizobium* clade in the *dnaK* trees, but within the *Agrobacterium* clades in both the 16S rRNA and DnaK trees, we compared the number of synonymous and nonsynonymous substitution differences between the *dnaK* genes of *R. galegae* and those of representatives of the *Agrobacterium* and *Rhizobium* clades (*A. rubi* and *R. tropici*, respectively). The results supported the differential placement of *R. galegae* between the *dnaK* and DnaK trees, in that (over the entire gene) the nucleotide sequence of *dnaK* in *R. galegae* has more synonymous site substitution differences from that of *A. rubi* than it does from that of *R. tropici* and, conversely, the *R. galegae* sequence has a slightly larger number of nonsynonymous substitution differences from the sequence of *R. tropici* than it does from the sequence of *A. rubi* (Table 3).

Polymorphic nucleotide position distribution across the 16S rRNA alleles. A map of the 182 polymorphic nucleotide positions present in the 16S rRNA sequences revealed some highly variable and some highly conserved regions (Fig. 3). The variable regions were correlated with expansion segments (e.g., V7) described by Raué et al. (15). Some obvious similarities and differences were evident among the polymorphic site patterns. For example, with the exception of the V7 region, the polymorphic site distribution patterns among the members of the *Agrobacterium* clade and *R. galegae* were similar (Fig. 3).

Analyses of mosaic structure within the 16S rRNA alleles. Evidence for recombination among the 16S rRNA alleles was obtained by the Stephens test. When applied without regard for specific partitions or groups, significant evidence of poly-

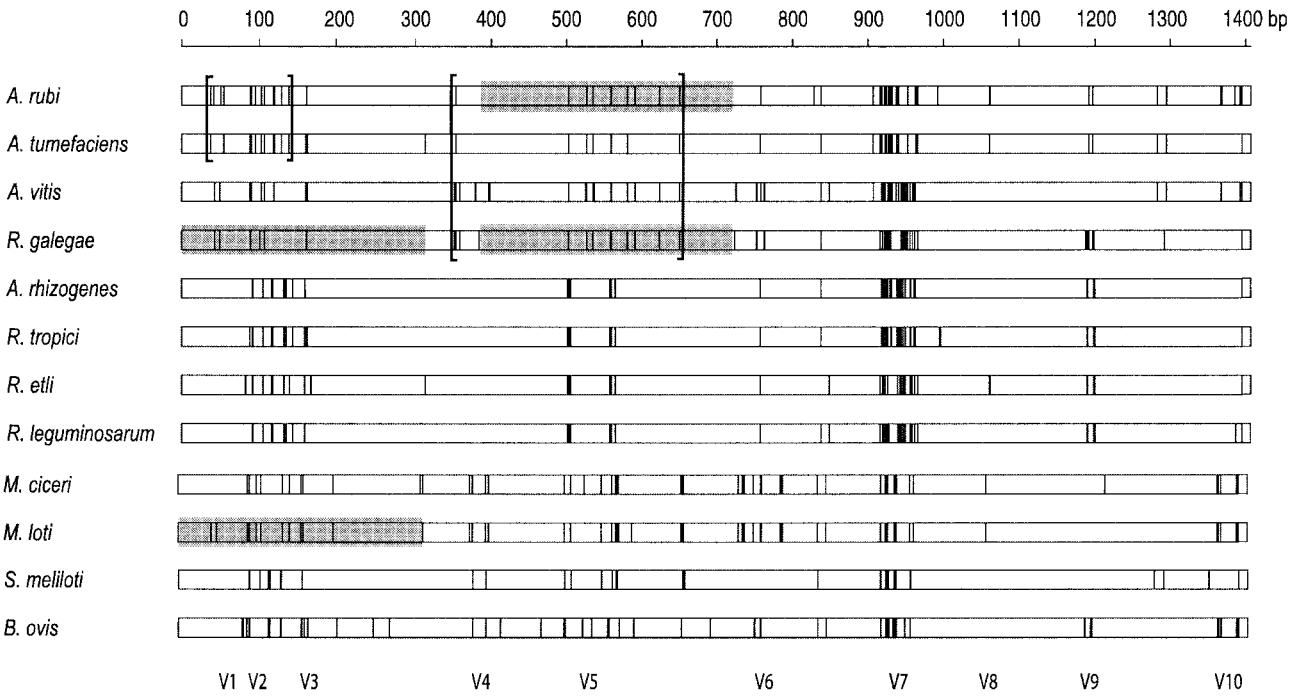

FIG. 3. Linear distribution of 182 polymorphic nucleotide positions in a multiple alignment of 16S rRNA sequences representing five genera (*Agrobacterium*, *Brucella*, *Mesorhizobium*, *Rhizobium*, and *Sinorhizobium*) within the order *Rhizobiales*. Each vertical line represents a deviation from the consensus sequence. The locations of hypervariable expansion segments (V1 through V10) described by Raué et al. (15) are shown at the bottom. Brackets define endpoints of segments that were identified by the Stephens test as containing nonrandom clusters of partition-specific nucleotide sequences. Shaded segments indicate runs of sequence similarity that were identified by Sawyer's Geneconv method.

morphic site clustering was apparent [$P\ (d \leq d_o) = 0.001$]. Because this method is sensitive to mutational hot (and cold) spots (see Materials and Methods), the highly polymorphic V7 region and the highly conserved region between V7 and V10 were deleted during the analysis. Regardless, significant partition-dependent clustering was observed within the 16S rRNA alleles, indicating that the recombinant segments represented by these clusters were not simply due to the presence of mutational hot (or cold) spots nearby. The locations of the sites supporting these clusters are enclosed by brackets in Fig. 3.

One of the partition-dependent clusters was located at the 5′ end of the 16S rRNA gene and spanned 102 bp in the alleles of *A. rubi* and *A. tumefaciens* [$P\ (d \leq d_o) = 0.001$]. The individual nucleotide positions supporting this partition were numbered 38′, 55′, 96′, 130′, and 140′ in the partial 16S rRNA alignment (Table 4). Only two alternative nucleotides were observed at these five positions, one of which was shared by only *A. rubi* and *A. tumefaciens*. Another cluster of partition-dependent polymorphic positions was located in a more central 292-bp segment, where *R. galegae* and three members of the *Agrobacterium* clade shared nucleotides at five positions. The nucleotides supporting this partition were located at positions 358, 529, 537, 582, and 650 [$P\ (d \leq d_o) = 0.023$] in the alignment (Fig. 3). Again, at each of these five positions only two alternative nucleotides were present, one of which was shared by only *R. galegae* and the three members of the *Agrobacterium* clade.

Supporting evidence for the distinctiveness of the 292-bp central segment (identified by the Stephens' test) was obtained with Sawyer's Geneconv program. A highly significant ($P \leq 0.001$) run of 41 identical polymorphic positions was identified in a 333-bp segment (spanning positions 389 through 721) in the sequences of *A. rubi* and *R. galegae*. These two identical segments are indicated in Fig. 3 by the shading in the central portion of the *A. rubi* and *R. galegae* distributions.

When the parameters in the Geneconv program were adjusted to relax the absolute sequence identity requirement, a second significant run of sequence similarity was identified at the 5′ end of the *R. galegae* and *M. loti* 16S rRNA genes—between nucleotide positions 1 and 318 ($P \leq 0.011$). Within this segment, there were a total of 43 polymorphic positions across the 12 sequences, 20 of which partitioned the sequence of *R. galegae* with either *M. loti* or *A. rubi*. Among these 20 positions, *R. galegae* shared specific nucleotides with *M. loti* at 15, while it shared only 5 with *A. rubi*. Seventeen of these positions were shown in Table 4. The remaining three were located downstream (not shown).

Because of the relatively high level of similarity between the *R. galegae* and *M. loti* sequences at the 5′ end of the gene and the absolute sequence identity between the *R. galegae* and *A. rubi* alleles in the more central 333-bp segment (identified by Sawyer's Geneconv), the degree of sequence similarity between the *R. galegae* and *M. loti* 16S rRNA alleles in the same 333-bp segment was determined. There were a total of 20 nucleotide mismatches between *R. galegae* and *M. loti* within the 333-bp segment, in contrast to the corresponding absolute sequence identity between the *A. rubi* and *R. galegae* alleles over the same segment.

Collectively these results indicated that the 5′ half of the *R.*

TABLE 4. Polymorphic nucleotide positions in the 16S rRNA genes (V1 through V3 region) of five genera in the order *Rhizobiales* (*Agrobacterium, Brucella, Mesorhizobium, Rhizobium,* and *Sinorhizobium*)

Species	16S rRNA nucleotide at position[a]:																										
	38′	42	43	50	51	55′	83	84	89	90	91	92	93	96′	102	104	106	108	117	119	120	121	130′	133	134	136	140′
A. rubi	A	T	C	G	A	T	T	C	C	A	A	C	**C**	G	T	G	T	T	G	A	A	T	C	A	C	**C**	A
A. tumefaciens	A	C	C	G	G	T	T	C	G	T	G	C	C	G	T	G	T	C	G	A	A	T	C	A	C	C	A
R. galegae	**G**	**C**	**T**	**A**	**G**	**C**	T	C	C	A	**T**	C	**C**	**A**	**C**	**A**	T	**C**	G	A	**G**	**C**	**T**	A	C	**C**	**T**
M. loti	**G**	**C**	**T**	**A**	**G**	**C**	T	C	C	A	**T**	C	T	**A**	**C**	**A**	T	**C**	G	A	**G**	**C**	**T**	A	C	T	**T**
A. vitis	G	C	T	A	G	C	T	C	G	T	A	C	C	A	T	G	T	C	G	A	A	T	T	A	C	C	T
A. rhizogenes	G	C	C	G	G	C	T	C	C	T	T	T	T	A	T	A	G	A	T	T	G	C	T	G	T	T	T
R. tropici	G	C	C	G	G	C	T	C	T	T	T	T	G	A	T	A	G	A	T	T	G	C	T	G	T	T	T
R. etli	G	C	C	G	G	C	C	G	C	T	T	T	A	A	T	A	G	A	T	T	G	C	T	G	T	C	T
R. leguminosarum	G	C	C	G	G	C	T	C	C	T	T	G	A	A	T	A	G	A	T	T	G	C	T	G	T	T	T
M. ciceri	G	C	C	G	G	C	T	C	C	A	T	C	T	A	C	A	T	C	G	A	G	C	T	A	C	T	T
S. meliloti	G	C	C	G	G	C	T	C	C	T	T	T	T	A	T	A	G	A	T	T	G	C	T	G	A	C	T
B. ovis	G	C	C	G	G	C	C	G	A	T	T	T	G	A	T	A	T	A	T	T	G	C	T	G	T	C	T

[a] Partition-specific nucleotide positions that were identified by the Stephens test are denoted by primes (38′, 55′, 96′, 130′, and 140′). Nucleotides shared between *R. galegae* and either *A. rubi* or *M. loti* are shown in boldface type.

galegae 16S rRNA gene can be divided into two distinct portions on the basis of clustered nucleotide sequence polymorphisms: one portion at the 5′ end of the gene, where the *R. galegae* sequence is more similar to that of *M. loti*, and a second more central portion, where the *R. galegae* sequence is identical to that of *A. rubi*. The presence of these two contrasting regions in the 16S rRNA gene of *R. galegae* could be explained either by independent parallel mutations at multiple sites or perhaps by the intragenic recombination of divergent 16S rRNA alleles.

If these segments of 16S rRNA in *R. galegae* indeed have independent evolutionary histories, then they would be expected to provide distinct phylogenetic signals. This assumption was examined by using the ILD test to determine whether alignment subsets corresponding to the 5′, the central, or the 3′ segments of the gene (described above) could be separated and then combined without significantly affecting the number of steps required for producing most parsimonious trees. The probabilities of congruence between the 5′ and remaining segments, the central segment and remaining segments, and also the 5′ and the central segment were $P = 0.002$, $P = 0.010$, and $P = 0.002$, respectively. It was therefore concluded that each of these segments produced different phylogenetic signals.

Polymorphic amino acid sequence position differences across the *dnaK* alleles. There were a total of 754 polymorphic nucleotide positions among the *dnaK* sequences (not shown). Because of the very large number of polymorphic positions, the graphical comparison method that was used for the 16S rRNA comparisons was visually uninformative for comparing the *dnaK* sequences. Consequently, the corresponding 177 polymorphic amino acid positions in the DnaK alignment were presented instead (Fig. 4). The highest degree of amino acid sequence polymorphism was evident at the carboxy-terminal end of the gene. In contrast, there was a strong consensus region at the 3′ end of the ATPase-binding domain, especially among the alleles of the *Rhizobium* clade (*A. rhizogenes*, *R. tropici*, *R. etli*, and *R. leguminosarum*). The polymorphic site distribution patterns of two *Mesorhizobium* species were the most similar, while the patterns of *S. meliloti* and *B. ovis* were the most different from each other and from the other species.

Analysis of mosaic structure within *dnaK* alleles. The Stephens test was used to search for evidence of intragenic recombination among the *dnaK* sequences. There was no significant evidence of clustering in the absence of specific phylogenetic partitions [$P\ (d \leq d_o) = 0.196$], thus justifying the use of the entire nucleotide alignment for the analysis of specific partitions. Among the 377 partitions examined, there was significant clustering in only 4; however, none of these could be explained by recombination or gene conversion. When applied to the DnaK amino acid sequence data, the Stephens test revealed no significant evidence of partition-specific clustering.

When Sawyer's Geneconv program was run with the *dnaK* nucleotide sequence data set, without an allowance for mismatches, no significantly long runs of similarity were detected. However, after relaxing the absolute sequence identity requirement, a 317-bp segment common to both *R. galegae* and *R. tropici* ($P \leq 0.039$) was revealed. This segment spanned nucleotide positions 862 through 1178 (corresponding to amino acid positions 288 through 393) in the *dnaK* sequences of *R. galegae* and *R. tropici* (Fig. 4). Across this segment, the two species differed at 14 of 121 polymorphic nucleotide positions (data not shown). The localized similarity between the *dnaK* segments of *R. galegae* and *R. tropici* at positions 862 through 1178 was supported by an analysis of the number of synonymous and nonsynonymous substitution differences within this segment (Table 3). The slightly lower number of nonsynonymous site differences between *R. galegae* and *R. tropici* in the segment from positions 862 through 1178, relative to the somewhat higher number observed between the corresponding *R. galegae* and *A. rubi* segments, did not reflect the pattern of nonsynonymous site differences that was observed between the same species over the entire *dnaK* gene. Although perhaps not statistically significant, these results support the distinctiveness of the segment from positions 862 through 1178.

DISCUSSION

A two-gene comparative analysis was used to reconstruct phylogenetic relationships among 12 species within the α-subdivision of the *Proteobacteria*. Nucleotide sequences for two conserved genes (16S rRNA and *dnaK*) were analyzed, and phylogenetic relatedness estimates were developed with these sequences (Fig. 1 and 2). Although certain groupings were

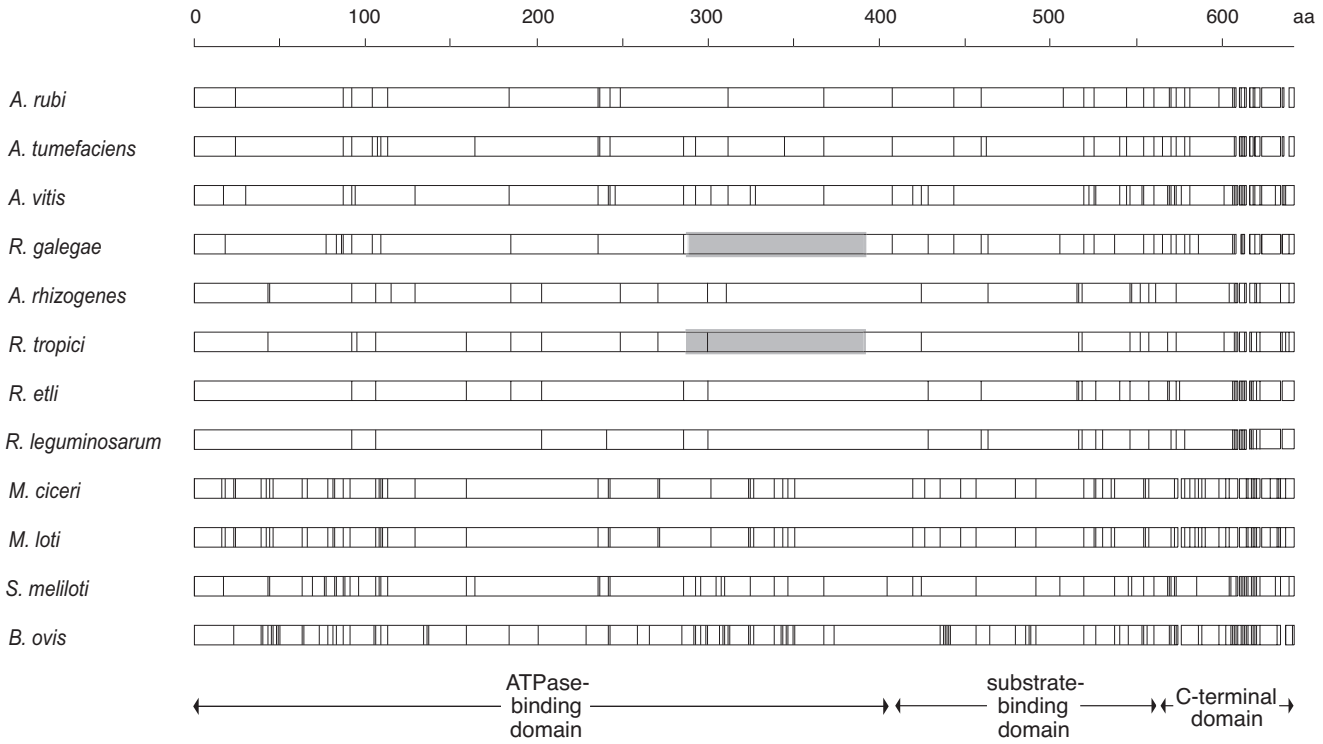

FIG. 4. Linear distribution of 193 polymorphic amino acid positions in a multiple alignment of DnaK protein sequences representing five genera (*Agrobacterium*, *Brucella*, *Mesorhizobium*, *Rhizobium*, and *Sinorhizobium*) within the order *Rhizobiales*. Each vertical line represents a deviation from the consensus sequence. Regions corresponding to the three domains described by Mogk et al. (13) are shown at the bottom. Shaded segments indicate statistically significant runs of nucleotide sequence similarity identified by Sawyer's method.

consistent between the trees representing the different loci (for example, the clustering of the two *Mesorhizobium* species and the clustering of *A. rhizogenes*, *R. tropici*, *R. etli*, and *R. leguminosarum*), some significant differences between the trees representing the different loci were evident. One of the most noticeable differences was the relative placement of the *R. galegae* sequences in the different trees, in that this species was placed flanking the *A. rhizogenes-Rhizobium* cluster in the *dnaK* nucleotide sequence trees, while it was placed along with the other three *Agrobacterium* species in both the 16S rRNA and DnaK amino acid trees. It was also noted that the bootstrap support for the placement of this particular species was modest in five of the six trees examined.

The lack of strong statistical support for the placement of *R. galegae* has been observed in other studies; however, conclusive explanations for this observation were not apparent. For example, in previous studies of rhizobial 16S rRNA genes, bootstrap support for the placement of this species has ranged from 54% (26) to 83% (24). Similarly, in a phylogenetic analysis of the 23S rRNA genes in rhizobia, Pulawska et al. (14) reported 59% bootstrap support for the placement of *R. galegae* with a representative of the *Agrobacterium* genus, *A. vitis*. In an analysis of glutamine synthetase I (GSI) sequences, Turner and Young (23) found no evidence supporting the placement of the *R. galegae* GSI allele with other GSI alleles of species representing the genus *Rhizobium*. However, in a parallel analysis of glutamine synthetase II (GSII) sequences, they reported strong support for the placement of the *R. galegae* allele with these same *Rhizobium* species. Unfortunately, no *Agrobacterium* sequences were included in their analyses, so the placement of the *R. galegae* glutamine synthetase alleles relative to those in *Agrobacterium* is unknown.

Differences or uncertainties in the phylogenetic estimates obtained through 16S rRNA sequence analyses are of concern because this locus is so widely used in applied studies for the identification of species and also in molecular systematics for the reconstruction of evolutionary relationships among bacteria (12). It is generally assumed that 16S rRNA genes are unlikely to have a history of lateral transfer and recombination because of the adverse effects that this would have on the translational efficiency within the cell. It is not difficult, however, to imagine a situation where recombinant 16S rRNA alleles might actually provide a selective advantage: for example, in the presence of species-specific rRNA-binding antibiotics. Evidence for this possibility was recently reported by Trieber and Taylor (22), who transformed tetracycline-sensitive *Helicobacter pylori* strains naturally to antibiotic resistance through the introduction of foreign 16S rRNA genes. Of course it is also possible, although probably less likely, that selectively neutral 16S rRNA alleles could be horizontally transferred and stably inherited in natural populations.

Several comparative studies of 16S rRNA gene diversity in bacteria have provided inferential evidence suggesting the existence of recombinant rRNA alleles in natural populations (4, 8, 19, 27, 29, 32). An extensive statistical analysis of rRNA genes in several species of the α-*Proteobacteria* has recently provided additional evidence suggesting a history of recombination between the 16S rRNA alleles of species of *Mesorhizo-*

bium and *Bradyrhizobium* and also between the 16S alleles of *Mesorhizobium* and *Sinorhizobium* (24).

To investigate the possibility that a history of recombination might explain the modest levels of bootstrap support observed for certain nodes within the trees in Fig. 1, we examined the sequence data used to generate those trees for patterns of clustered nucleotide sequence substitutions. Perhaps the strongest evidence for the mosaic gene structure that we observed was the presence of clustered substitutions between the 16S rRNA alleles of *A. rubi* and *R. galegae* (Fig. 3). While both of these species shared an identical 333-bp segment spanning the hypervariable V4-V5 regions of their 16S rRNA genes, both had distinctly different nucleotide substitution patterns upstream of this segment. In this upstream region (V1 through V3), most of the polymorphic nucleotides in *A. rubi* matched those of *A. tumefaciens*, while most of those in *R. galegae* matched those of *M. loti* (Table 4). A rational explanation for these results is that different regions of the *R. galegae* 16S rRNA gene have different evolutionary histories.

There are, however, other plausible explanations for the significant runs of sequence similarity observed among the 16S rRNA and *dnaK* nucleotide sequences of these species. For example, rates of evolution may differ in different regions of a particular gene (28). Selection for compensatory mutations might also help to stabilize secondary structures of certain gene products: e.g., stem structures in rRNA. This could also influence the relative frequency of compatible positions in the sequence alignments. Regardless of the specific mechanism responsible for the clustered patterns of nucleotide substitutions that we observed, the results of this study are important in that they demonstrate how such patterns can be used to explain topological instabilities that are often encountered in sequence-based phylogenetic trees. The results are also of applied significance in that they expose a potential pitfall in the use of partial gene sequences as a rapid method for species identification.

ACKNOWLEDGMENTS

This research was supported by Public Health Service grant AI 22144 (R.K.S.) and a Penn State Berks-Lehigh Valley College Research and Development grant (B.D.E.).

The authors gratefully acknowledge T. S. Whittam for providing software and advice on statistical analyses.

REFERENCES

1. **Boorstein, W. R., T. Zeigelhoffer, and E. A. Craig.** 1994. Molecular evolution of the HSP70 multigene family. J. Mol. Evol. **38:**1–17.
2. **Brosius, J., T. J. Dull, D. N. Sleeter, and H. F. Noller.** 1981. Gene organization and primary structure of a ribosomal RNA operon from *Escherichia coli*. J. Mol. Biol. **148:**107–127.
3. **Cellier, M. F. M., J. Teyssier, M. Nicolas, J. P. Liautard, J. Marti, and J. Sri Widada.** 1992. Cloning and characterization of the *Brucella ovis* heat shock protein DnaK functionally expressed in *Escherichia coli*. J. Bacteriol. **174:** 8036–8042.
4. **Eardly, B. D., F.-S. Wang, and P. van Berkum.** 1996. Corresponding *16S* rRNA gene segments in *Rhizobiaceae* and *Aeromonas* yield discordant phylogenies. Plant Soil **186:**69–74.
5. **Falah, M., and R. S. Gupta.** 1994. Cloning of the *hsp70* (*dnaK*) genes from *Rhizobium meliloti* and *Pseudomonas cepacia*: phylogenetic analyses of mitochondrial origin based on a highly conserved protein sequence. J. Bacteriol. **176:**7748–7753.
6. **Farrand, S. K., P. van Berkum, and P. Oger.** 2003. *Agrobacterium* is a definable genus of the family *Rhizobiaceae*. Int. J. Syst. Evol. Microbiol. **53:**1681–1687.
7. **Farris, J. S., M. Källersjö, A. G. Kluge, and C. Bult.** 1994. Testing significance of incongruence. Cladistics **10:**315–319.
8. **Gaunt, M. W., S. L. Turner, L. Rigottier-Gois, S. A. Macgilp, and J. P. W. Young.** 2001. Phylogenies of *atpD* and *recA* support the small subunit rRNA-based classification of rhizobia. Int. J. Syst. Evol. Microbiol. **51:**2037–2048.
9. **Gething, M.-J., and J. Sambrook.** 1992. Protein folding in the cell. Nature **355:**33–45.
10. **Hartl, F. U.** 1996. Molecular chaperones in cellular protein folding. Nature **381:**571–580.
11. **Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei.** 2001. MEGA 2: molecular evolutionary genetic analysis software. Bioinformatics **17:**1244–1245.
12. **Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, Jr., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje.** 2001. The RDP-II (Ribosomal Database Project). Nucleic Acids Res. **29:**173–174.
13. **Mogk, A., B. Bukau, R. Lutz, and W. Schumann.** 1999. Construction and analysis of hybrid *Escherichia coli-Bacillus subtilis dnaK* genes. J. Bacteriol. **181:**1971–1974.
14. **Pulawska, J., M. Maes, A. Willems, and P. Sobiczewski.** 2000. Phylogenetic analysis of 23S rRNA gene sequences of *Agrobacterium*, *Rhizobium*, and *Sinorhizobium* strains. Syst. Appl. Microbiol. **23:**238–244.
15. **Raué, H. A., W. Musters, C. A. Rutgers, J. Van't Riet, and R. J. Planta.** 1990. rRNA: from structure to function, p. 217–235, *In* W. E. Hill, P. B. Moore, A. Dahlberg, D. Schlessinger, R. A. Garret, and J. R. Warner (ed.), The ribosome: structure, function, and evolution. American Society for Microbiology, Washington, D.C.
16. **Sambrook, J. E., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
17. **Sawyer, S. A.** 1999. GENECONV: a computer package for the statistical detection of gene conversion., 1.02 ed. Department of Mathematics, Washington University, St. Louis, Mo.
18. **Segal, G., and E. Z. Ron.** 1995. The *dnaKJ* operon of *Agrobacterium tumefaciens*: transcriptional analysis and evidence for a new heat shock promoter. J. Bacteriol. **177:**5952–5958.
19. **Sneath, P. H. A.** 1993. Evidence from *Aeromonas* for genetic crossing-over in ribosomal sequences. Int. J. Syst. Bacteriol. **43:**626–629. (Letter.)
20. **Stephens, J. C.** 1985. Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:**539–556.
21. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 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.
22. **Trieber, C. A., and D. E. Taylor.** 2002. Mutations in the 16S rRNA genes of *Helicobacter pylori* mediate resistance to tetracycline. J. Bacteriol. **184:**2131–2140.
23. **Turner, S. L., and J. P. W. Young.** 2000. The glutamine synthetases of rhizobia: phylogenetics and evolutionary implications. Mol. Biol. Evol. **17:** 309–319.
24. **van Berkum, P., Z. Terefework, L. Paulin, S. Suomalainen, K. Lindström, and B. D. Eardly.** 2003. Discordant phylogenies within the *rrn* loci of rhizobia. J. Bacteriol. **185:**2988–2998.
25. **Vincent, J. M.** 1970. A manual for the practical study of root nodule bacteria. IBP handbook, no. 15. Blackwell Scientific Publications, Ltd., Oxford, England.
26. **Wang, E. T., P. van Berkum, D. Beyene, X. H. Sui, O. Dorado, W. X. Chen, and E. Martinez-Romero.** 1998. *Rhizobium huautlense* sp. nov., a symbiont of *Sesbania herbacea* that has a close phylogenetic relationship with *Rhizobium galegae*. Int. J. Syst. Bacteriol. **48:**687–699.
27. **Wang, Y., Z. Zhang, and N. Ramanan.** 1997. The actinomycete *Thermobispora bispora* contains two distinct types of transcriptionally active 16S rRNA genes. J. Bacteriol. **179:**3270–3276.
28. **Worobey, M., A. Rambaut, O. G. Pybus, and D. L. Robertson.** 2002. Questioning the evidence for genetic recombination in the 1918 "Spanish Flu" virus. Science **296:**211–213.
29. **Yap, W. H., Z. Zhang, and Y. Wang.** 1999. Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J. Bacteriol. **181:**5201–5209.
30. **Young, J. M., L. D. Kuykendall, E. Martinez-Romero, A. Kerr, and H. Sawada.** 2001. A revision of *Rhizobium* Frank 1889, with an emended description of the genus, and the inclusion of all species of *Agrobacterium* Conn 1942 and *Allorhizobium undicola* de Lajudie et al. 1998 as new combinations: *Rhizobium radiobacter*, *R. rhizogenes*, *R. rubi*, *R. undicola* and *R. vitis*. Int. J. Syst. Evol. Microbiol. **51:**89–103.
31. **Young, J. M., L. D. Kuykendall, E. Martinez-Romero, A. Kerr, and H. Sawada.** 2003. Classification and nomenclature of *Agrobacterium* and *Rhizobium*. Int. J. Syst. Evol. Microbiol. **53:**1689–1695.
32. **Young, J. P. W., and K. E. Haukka.** 1996. Diversity and phylogeny of rhizobia. New Phytol. **133:**87–94.