Phylogenetic relationships and genotyping of the genus *Streptococcus* by sequence determination of the RNase P RNA gene, *rnpB*

Jenny Täpp,[1] Mikael Thollesson[2] and Björn Herrmann[1]

[1]Department of Clinical Microbiology, University Hospital, SE-751 85 Uppsala, Sweden

[2]Department of Molecular Evolution, EBC, Uppsala University, Norbyvägen 18C, SE-195 30 Uppsala, Sweden

Correspondence
Björn Herrmann
bjorn.herrmann@medsci.uu.se

The *rnpB* gene is universally present in bacterial species and encodes the RNA subunit of endoribonuclease P. In this study, *rnpB* was sequenced in 50 type strains and 29 additional strains of the genus *Streptococcus*. Putative secondary-structure models and possible interactions in RNase P RNA molecules are discussed. Phylogenetic relationships were studied and Bayesian, maximum-parsimony and minimum-evolution analyses supported six main clades that comprised 22 of the 50 species analysed. Phylogenetic inference was also studied for the 16S rRNA gene; it indicated a similar tree topology, but with weaker support values than for *rnpB*. Combined analysis of *rnpB* and 16S resulted in a phylogeny with significantly better support. Variability in the *rnpB* and 16S genes among all type strains, calculated as Shannon–Wiener information index values, was 0·45 for *rnpB* and 0·15 for 16S. Intraspecies proximity was assessed by principal coordinate analysis of *rnpB* for 32 strains of six closely related species (two clades) and showed species-specific clusters, but heterogeneity occurred in two species. It can be concluded that the *rnpB* gene is suitable for phylogenetic analysis of closely related taxa and has potential as a tool for species discrimination.

INTRODUCTION

Identification of *Streptococcus* species has traditionally been based on the serological grouping by Lancefield and haemolytic reactions. However, Lancefield groups are not species-specific (Farrow & Collins, 1984; Lawrence *et al.*, 1985) and haemolytic activity differs within species and depends on incubation procedures, as well as origin of blood in the substrate. In clinical laboratories, current means of identification of streptococci rely on phenotypic tests. These largely provide adequate species differentiation, but up to 13 % of analysed strains may be identified incorrectly (Kikuchi *et al.*, 1995). Furthermore, strains within a given species may differ for a common trait (Kilian *et al.*, 1989; Beighton *et al.*, 1991) and the same strain may exhibit biochemical variability (Hillmann *et al.*, 1989; Tardif *et al.*, 1989). Moreover, small alterations in realization of a phenotypic test may give false results.

Although no single classification system is perfect, the development of genetic analysis has improved the identification of streptococci. DNA hybridization (Schmidhuber *et al.*, 1988; Adnan *et al.*, 1993; Bentley & Leigh, 1995; Jacobs *et al.*, 1996), rDNA restriction analysis (Jayarao *et al.*, 1992; Rudney & Larson, 1994; Gillespie *et al.*, 1997), amplification of selected targets [interspace 16S–23S (Saruta *et al.*, 1995; Whiley *et al.*, 1995), *ddl* gene (Garnier *et al.*, 1997), tDNA PCR fragment length (De Gheldre *et al.*, 1999; Baele *et al.*, 2001)] and sequence analysis (Bentley *et al.*, 1991; Poyart *et al.*, 1998) are alternative techniques that provide differentiation between species, often with high resolution.

Differentiation of closely related chlamydial species has recently been shown to be possible by using the RNase P RNA gene *rnpB* as a marker (Herrmann *et al.*, 1996, 2000). Endoribonuclease P (RNase P) is a ribonucleoprotein complex that removes 5′ leader sequences from tRNA precursors during tRNA biosynthesis. RNase P is present in all cells and subcellular compartments that synthesize tRNA, but catalytic activity by RNA alone has only been demonstrated for bacterial and some archaeal RNase P RNAs (Pannucci *et al.*, 1999). The endoribonuclease has been characterized best in the division *Bacteria*, where it is composed of an RNA molecule of approximately 400 nt and a protein of about 120 aa (Altman & Kirsebom, 1999). Bacterial RNase P RNAs have been separated into two main classes based on secondary structure: type A is the most common class, whereas type B is found exclusively in

Published online ahead of print on 16 May 2003 as DOI 10.1099/ijs.0.02639-0.

Abbreviations: ME, minimum-evolution; MP, maximum-parsimony.

The GenBank/EMBL/DDBJ accession numbers for *rnpB* are listed in Table 1.

the low G+C Gram-positive *Bacteria* (Haas *et al.*, 1996). Secondary structure of RNase P RNA has been characterized for many bacterial lineages and variation among the helices provides useful phylogenetic information (Haas & Brown, 1998).

In this study, we sequenced the *rnpB* gene in 50 *Streptococcus* species, performed phylogenetic analysis and evaluated the possibility of using *rnpB* for genotyping.

METHODS

Bacterial strains. The 79 streptococcal strains (50 type strains and 29 other strains) used in this study and their sources are listed in Table 1.

DNA preparations. DNA was extracted from solubilized freeze-dried strains or from cultured material by using a QIAamp DNA Mini kit (Qiagen) according to the manufacturer's instructions.

PCR amplification and sequencing. The *rnpB* gene from *Streptococcus* species was amplified by PCR with the primer pair strF (5′-YGTGCAATTTTTGGATAAT-3′) and strR (5′-TTCTATAA-GCCATGTTTTGT-3′); the design was based on the *Streptococcus pyogenes* sequence (Brown, 1999). The reaction mixture contained 0·2 μM each primer, 200 μM dNTPs, 2·5 mM $MgCl_2$ and 1·5 U HotStarTaq DNA polymerase (Qiagen). Amplification conditions consisted of 15 min enzyme activation at 95 °C, followed by five cycles of 40 s at 94 °C, 40 s at 58 °C and 40 s at 72 °C. An additional five cycles were added with an annealing temperature of 54 °C, followed by 30 cycles with an annealing temperature of 50 °C. To generate a PCR product from *Streptococcus pleomorphus*, primers from conserved *rnpB* regions were used: BAC1 [5′-(GCT)NRG-(GCT)NGAGGAAAGTCC-3′] and BM1-2 (5′-TGTAAAACGGCCA-GTRTAAGCCGGGTTCTGT-3′). Amplification conditions were as above, except that the primer concentration was 0·4 μM and the annealing temperature was 43 °C for all 40 cycles.

PCR products were sequenced by using BigDye Terminator labelled cycle sequencing chemistry (Applied Biosystems) and sequence reactions from complementary strands were analysed on a 310 Genetic Analyser (Applied Biosystems). Consensus sequences were submitted to GenBank/EMBL and all accession numbers of *rnpB* sequences from the present study are listed in Table 1.

Phylogenetic analysis. Sequence alignment required secondary-structure modelling of RNase P RNA, which was performed manually by using comparative sequence analysis. Sequences of 16S rRNA were obtained from the Ribosomal Database Project (Maidak *et al.*, 2001) and GenBank. Obtained alignments were used for phylogenetic inference by using a Bayesian approach, as implemented in MRBAYES 2.01 (Huelsenbeck & Ronquist, 2001). MRBAYES uses Metropolis-coupled Markov chain Monte Carlo methods to calculate posterior probabilities for parameters of interest. Each analysis was run for 5×10^5 generations and the first 5×10^4 were discarded as burn-in; four differently heated chains were employed. Convergence was tested by rerunning the analyses with different initial parameter values.

To select an adequate model for the Bayesian analysis (as well as pairwise distances), we used a hierarchical likelihood ratio test (ηLRT) approach (Huelsenbeck & Crandall, 1997). To do this, we used PAUP* (4.0b8–10, Linux and Macintosh versions; Swofford, 2000) and the same test hierarchy (and thus model selection) as implemented in the program MODELTEST (Posada & Crandall, 1998) at $P<0{\cdot}01$. Neighbour-joining trees under the Jukes–Cantor model were produced for each of the datasets/partitions, separately as well as combined, and parameters for each model were estimated by using these trees. The same model was used for pairwise distances (e.g. γ-shape and proportion of invariant sites) and the parameters assigned were based on the maximum-likelihood estimate.

In addition to the Bayesian analysis, we performed bootstrap analyses by using maximum-parsimony (MP) and minimum-evolution (ME) as optimality criteria, with PAUP* 4·0b8–10. For the optimality criterion methods, 1000 bootstrap replicates were performed and heuristic search algorithms were used, namely simple stepwise addition and TBR (tree bisection–reconnection) branch-swapping.

To map the hosts of the various species on the phylogenetic hypothesis, we used MacClade 3.08 on the combined dataset of 16S and *rnpB*.

To compare variation in the different genes, we used the Shannon–Wiener information index, H, defined as:

$$H=-\sum_{i=1}^{4}p_i\log p_i$$

where p_i is the proportion of A, T, C and G (Shannon & Weaver, 1949; Wiener, 1949). The mean value for all sites in the same set of taxa was calculated for each dataset.

To show similarity and distinctiveness of closely related species, we did principal coordinate analysis (Gower, 1966) on uncorrected pairwise distances (p-distances) by using the software DISTPCOA (Legendre & Anderson, 1998, 1999).

RESULTS AND DISCUSSION

Comparison of *rnpB* sequences and characterization of secondary-structure models

PCR products that included 97·5 % of the full-length *rnpB* gene were obtained from 79 streptococcal strains, representing 49 different species. Products varied in length from 367 to 388 bp and similarities were 77·7–100 %. Additionally, a 304 bp product was obtained from *S. pleomorphus*, making a total of 50 species examined. The only species with identical sequences were *Streptococcus waius* and *Streptococcus macedonicus*, both isolated originally from dairy products and proposed recently to comprise a single species (Manachini *et al.*, 2002; Poyart *et al.*, 2002). Alignment of *rnpB* sequences indicated seven hypervariable regions located in distinct stem–loops, denoted P3, P9, P10.1, P12, P15.1 and P19 in the suggested secondary structure (Fig. 1). This is in keeping with previous reports (Haas *et al.*, 1996; Massire *et al.*, 1998); it is thought that these hypervariable domains of the RNA subunit of RNase P RNA are important in the formation of a mature ribozyme and its interaction with tRNA precursors.

The RNase P RNA molecule is, unlike rRNA, quite variable in secondary structure outside the conserved core; therefore, previous studies have focused on structural comparisons in different bacterial genera (Haas & Brown, 1998; Massire *et al.*, 1998; and references therein). In the present study, *rnpB* was compared in 50 species within a single genus.

The P15.2 helix varies in length among the species

Table 1. Strains, GenBank accession numbers of *rnpB* and 16S rRNA gene sequences and sources of strains used

ATCC, American Type Culture Collection, Manassas, VA, USA; CCUG, Culture Collection University of Göteborg, Gothenburg, Sweden; Claesson, R. Claesson, Department of Odontology, Section of Oral Microbiology, Umeå University, Umeå, Sweden; Leigh, J. Leigh, Institute for Animal Health, Compton, Berkshire, UK; SMI, Swedish Institute for Infectious Disease Control, Stockholm, Sweden.

Strain	GenBank accession no.		Source
	rnpB	16S rRNA*	
S. acidominimus ATCC 51725^T = CCUG 27296^T	AJ511681	X58301	CCUG
S. agalactiae ATCC 13813^T = CCUG 4208^T	AJ511673	AB002479	CCUG
S. alactolyticus ATCC 43077^T = CCUG 27297^T	AJ511706	X58319	CCUG
S. anginosus CCUG 27298^T	AJ511744	X58309	CCUG
S. anginosus CCUG 35271	AJ511736		SMI
S. anginosus CCUG 27619	AJ511735		CCUG
S. anginosus CCUG 28191	AJ511731		CCUG
S. anginosus CCUG 28192	AJ511732		CCUG
S. anginosus CCUG 44103	AJ511739		CCUG
S. anginosus CCUG 44890	AJ511733		CCUG
S. anginosus CCUG 44758	AJ511740		CCUG
S. bovis ATCC 33317^T = CCUG 17828^T†	AJ511685	X58317	Leigh
S. bovis ATCC 33317 = CCUG 34832	AJ511711		CCUG
S. bovis CCUG 4214	AJ511745		CCUG
S. canis ATCC 43496^T = CCUG 27661^T	AJ511684	AB002483	CCUG
S. constellatus ATCC 27823^T = CCUG 24889^T	AJ511728	Z69041	SMI
S. constellatus UMU KE	AJ511727		Claesson
S. constellatus CCUG 4215	AJ511741		SMI
S. constellatus CCUG 9569	AJ511730		CCUG
S. constellatus CCUG 28196	AJ511742		CCUG
S. criceti ATCC 19642^T = CCUG 27300^T	AJ511689	X58305	CCUG
S. cristatus ATCC 51100^T = CCUG 33481^T	AJ511700	AB008313	SMI
S. downei ATCC 33748^T = CCUG 24890^T	AJ511699	X58306	CCUG
S. dysgalactiae subsp. *dysgalactiae* ATCC 43078^T = CCUG 27301^T	AJ511668	AB002485	CCUG
S. dysgalactiae subsp. *equisimilis* CCUG 36637^T	AJ512494	AB008926	CCUG
S. equi subsp. *equi* ATCC 33398^T = CCUG 23255^T	AJ511695	AB002515	CCUG
S. equi subsp. *zooepidemicus* ATCC 43079^T = CCUG 23256 ^{T}	AJ511679	AB002516	SMI
S. equinus ATCC 9812^T = CCUG 27302^T	AJ511669	AJ301607	CCUG
S. ferus ATCC 33477^T = CCUG 34834^T	AJ511705	AY058218	CCUG
S. gallolyticus CCUG 35224^T	AJ511683	X94337	CCUG
S. gordonii ATCC 10558^T = CCUG 33482^T	AJ511701	AF003931	CCUG
S. hyointestinalis ATCC 49169^T = CCUG 27888^T	AJ511696	X58313	Leigh
S. hyovaginalis CCUG 37866^T	AJ512493	Y07601	CCUG
S. infantarius subsp. *coli* ATCC BAA-103^T = CCUG 43822^T (*S. lutetiensis*)	AJ511709	AF177729	CCUG
S. infantarius subsp. *infantarius* ATCC BAA-102^T = CCUG 43820^T	AJ511688	AF429762	CCUG
S. infantis ATCC 700779^T = CCUG 39817^T	AJ511687	AB008315	CCUG
S. iniae ATCC 29178^T = CCUG 27303^T	AJ511708	AF335572	CCUG
S. intermedius ATCC 27335^T = CCUG 32759^T	AJ511734	AF104671	CCUG
S. intermedius ATCC 27335 = CCUG 17827‡	AJ511729		SMI
S. intermedius CCUG 28203	AJ511737		CCUG
S. intermedius CCUG 28204	AJ511738		CCUG
S. intermedius UMU 90A	AJ511743		Claesson
S. macacae ATCC 35911^T = CCUG 27653^T	AJ511702	X58302	CCUG
S. macedonicus ATCC BAA-249^T = CCUG 39970^T	AJ511677	Z94012	CCUG
S. mitis ATCC 49456^T = CCUG 31611^T	AJ511694	D38482	CCUG
S. mutans ATCC 25175^T = CCUG 6519^T	AJ511678	U02919	SMI
S. oralis ATCC 35037^T = CCUG 13229^T	AJ511698	AF003932	CCUG
S. orisratti ATCC 700640^T = CCUG 43577^T	AJ511692	AF124350	CCUG
S. parasanguinis ATCC 15912^T = CCUG 30417^T	AJ511704	AF003933	CCUG

Table 1. cont.

Strain	GenBank accession no.		Source
	rnpB	16S rRNA*	
S. parauberis CCUG 39954^T	AJ511676	X89967	CCUG
S. peroris ATCC 700780^T = CCUG 39814^T	AJ511690	AB008314	CCUG
S. phocae ATCC 51973^T = CCUG 35103^T	AJ511670	AF235052	CCUG
S. pleomorphus ATCC 29734^T = CCUG 11733^T	AJ511710	M23730	CCUG
S. pluranimalium ATCC 700864^T = CCUG 43803^T	AJ511697	Y18026	CCUG
S. pneumoniae ATCC 33400^T = CCUG 28588^T	AJ511703	X58312	SMI
S. porcinus ATCC 43138^T = CCUG 27628^T	AJ511675	AB002523	CCUG
S. pyogenes ATCC 12344^T = CCUG 4207^T	AJ511686	AB002521	SMI
S. ratti ATCC 19645^T = CCUG 27642^T	AJ511671	X58304	CCUG
S. salivarius ATCC 7073^T = CCUG 11878^T	AJ511715	X58320	CCUG
S. salivarius CCUG 33081B	AJ511722		CCUG
S. salivarius CCUG 32718	AJ511726		CCUG
S. salivarius CCUG 41462	AJ511716		CCUG
S. salivarius ATCC 13419 = CCUG 33776	AJ511723		CCUG
S. sanguinis ATCC 10556^T = CCUG 17826^T	AJ511682	AF003928	CCUG
S. sobrinus ATCC 33478^T = CCUG 25735^T	AJ511707	AJ243966	CCUG
S. suis ATCC 43765^T = CCUG 7984^T	AJ511674	AF009477	CCUG
S. thermophilus ATCC 19258^T = CCUG 21957^T	AJ511725	X68418	CCUG
S. thermophilus CCUG 30577	AJ511717		CCUG
S. thermophilus CCUG 35458	AJ511712		CCUG
S. thermophilus CCUG 43039	AJ511719		CCUG
S. thermophilus CCUG 43382	AJ511718		CCUG
S. uberis ATCC 19436^T = CCUG 17930^T	AJ511693	AB002526	CCUG
S. urinalis CCUG 41590^T	AJ511680	AJ131965	CCUG
S. vestibularis ATCC 49124^T = CCUG 24893^T	AJ511721	X58321	CCUG
S. vestibularis CCUG 24683	AJ511724		CCUG
S. vestibularis CCUG 24684	AJ511720		CCUG
S. vestibularis CCUG 41623	AJ511713		CCUG
S. vestibularis CCUG 45661	AJ511714		CCUG
S. waius CCUG 43003^T	AJ511672	AF088900	CCUG

*Accession numbers for 16S rRNA sequences refer to the same type strain as that for *rnpB* sequences. Exceptions are: *S. acidominimus* X58301; *S. infantarius* subsp. *coli* (*S. lutetiensis*) AF177729 strain HDP90104, not denoted as type strain; *S. parauberis* X89967, no strain designation given in GenBank; *S. pluranimalium* strain LMG 14257, not denoted as type strain.

†Strains CCUG 17828 and CCUG 34832 are equivalent strains with the same ATCC designation, but different histories. The *rnpB* sequence of *S. bovis* in GenBank is represented by an undefined strain; this sequence (accession number AF295988) has 23 positions with discordant nucleotides compared to CCUG 17828^T (AJ511685) and five discordant nucleotides compared to *S. gallolyticus* CCUG 35224^T (AJ511683).

‡Strains CCUG 32759 and CCUG 17827 are equivalent strains with the same ATCC designation, but different histories.

examined. In five species that belong to the mitis group, this helix was 29 nt long, whereas in all other species it was 12–16 nt long. This is in keeping with the tertiary model of RNase P RNA, in which this helix can exhibit variable length in type B sequences without affecting the interaction of the P15.1/P15.2 helix with the loop of the P5 helix (Massire *et al.*, 1998).

The P15.1 loop has been suggested to interact with the P5.1 loop to stabilize the catalytic site of RNase P RNA. Conserved motifs of RAA-NNNAA in P15.1 and UGNRAU in P5.1 would participate in this interaction. Sequences from streptococci fitted into this model and the corresponding sequence variation in the motifs was GAA-N(C/A/U)GAA and UG(T/A/C)GAU, respectively.

Another tertiary interaction has been proposed for the distal part of P10.1 and a highly conserved GAAA tetraloop of P12 (Tanner & Cech, 1995; Haas *et al.*, 1996). Among streptococci, the GAAA loop was found in all sequences except for that of *Streptococcus bovis*, which had AAAA in this loop. This observation in the type strain was confirmed in sequence analysis of two other *S. bovis* strains (CCUG 34832 and CCUG 4214). A corresponding shift in the motif of the P10.1 stem was not detected. RNase P RNA of type B typically forms an internal loop in P10.1 (nt 136–140 and

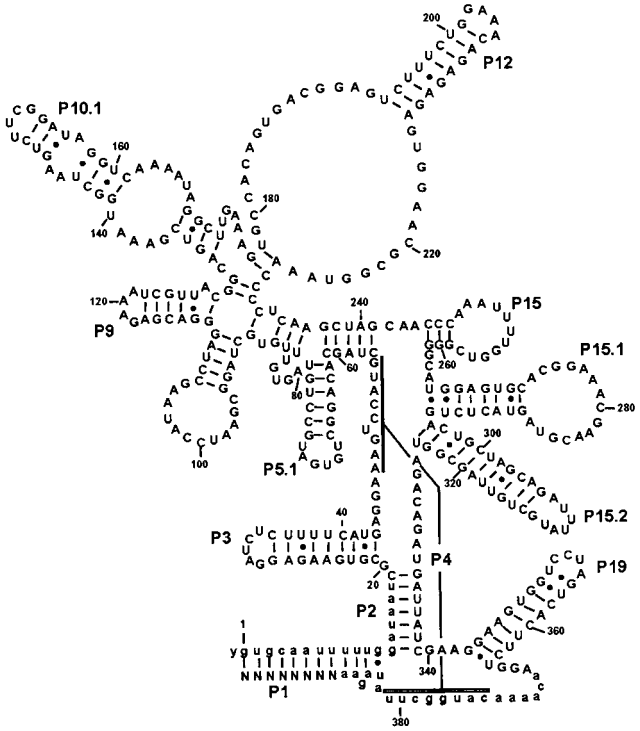

Fig. 1. Deduced secondary structure of RNase P RNA in *Streptococcus oralis*. Nucleotides in primer sequences are in lower case. Abbreviations: y, cytosine/thymidine mixture; N, tentative nucleotides in flanking regions of primer, based on minimum consensus bacterial RNase P RNA (Brown, 1999). Suggested long-distance interactions for P4 are denoted by lines.

162–167 in Fig. 1, *Streptococcus oralis*) with the consensus motif RAA...RAGUA (Fig. 3 of Massire *et al.*, 1998). Of 49 streptococcal type strains, the RAAA motif was not found in nine strains. These strains differed by 1 or 2 nt. The RAGUA motif was found only in 19 species and of the 30 species with alterations, three species had three nucleotide changes (two with two gaps) and one species had changes in all five positions. Variation noted in the internal P10.1 loop shows the complexity of the helix and the difficulty in interpreting structure patterns.

We found that the P3 helix varied in length between 15 and 27 nt. The P19 helix was almost invariable in length (21–24 nt), but varied in the number of nucleotides forming the suggested loop (4–10 nt without canonical base-pairing or wobble pairs). These two findings are compatible with the proposed three-dimensional structure model of the RNA molecule, in which P19, P2 and P3 form a continuous helical stack of the domain responsible for catalytic activity (Massire *et al.*, 1998). Observed variation in P3 and P19 occurs at the distal ends of the P19/P2/P3 stack and is therefore less likely to affect the three-dimensional model.

Analysis of a 304 bp *rnpB* fragment of *S. pleomorphus*, a species suggested to be removed from the *Streptococcus* genus (Ludwig *et al.*, 1988), showed some distinctive features compared to those of classical streptococcal species. The P10.1 loop formed a closed stem of 13 bp without the internal loop that is typical of type B RNase P RNA. Furthermore *S. pleomorphus* completely lacks the P19 stem–loop structure, which is found in all other streptococcal species. Our findings are compatible with the proposed removal of *S. pleomorphus* from the genus *Streptococcus*.

Phylogeny of the genus *Streptococcus*

Different approaches were used to infer phylogenetic relationships among streptococcal species. Analysis of both the *rnpB* and 16S rRNA genes required that site-to-site rate variation was modelled; the optimal model for both genes used a discrete γ-distribution (Yang, 1993, 1994) and treated a fraction of the sites as invariant (Gu *et al.*, 1995; Waddell & Penny, 1996). For *rnpB*, as well as the combined dataset of *rnpB* and 16S, the Tamura–Nei model (i.e. unequal base frequencies, transversions and two classes of transitions treated separately; Tamura & Nei, 1993) was indicated to be the most adequate, whereas the 16S dataset required the general time-reversible (GTR) model with six different rate parameters (Rodriguez *et al.*, 1990). Due to limitations in the MRBAYES software, the model actually employed for Bayesian analysis of the combined dataset was the more parameter-rich GTR model, with separate γ-distributions for the genes.

To compare phylogenetic utility of the two genes, and specifically to explore clade support in each gene and any significant conflicts between them, majority-rule consensus trees that comprise clades with a posterior probability of $\geqslant 0{\cdot}85$ for the two genes when analysed separately, are shown in Fig. 2. Branch support, as assessed by Bayesian posterior probabilities and bootstrap percentages in the MP and ME analyses, are given on the branches. Results from the two genes were generally congruent, i.e. there were only a few cases where clades with substantial support were in conflict (see below) and all remaining differences in the optimal trees (not shown) from the two genes can, at the moment, be treated as being due to limited sampling (i.e. sequence length) and not as real incongruence between the histories of the two genes. Thus, a combined approach (using both genes in the analysis) is beneficial to obtain a better phylogenetic hypothesis of the genus *Streptococcus* (Fig. 3).

Previously, a study of the 16S gene in 34 type strains of streptococci separated the genus into at least six clusters (Kawamura *et al.*, 1995), albeit with no information on how well-supported by the data these clusters were (e.g. as evaluated by bootstrap). That classification was based only on the neighbour-joining method and further details of the construction of the unrooted tree were not mentioned. In our analysis of the 16S rRNA gene, four main clades with significant support (i.e. posterior probability of $\geqslant 0{\cdot}95$)

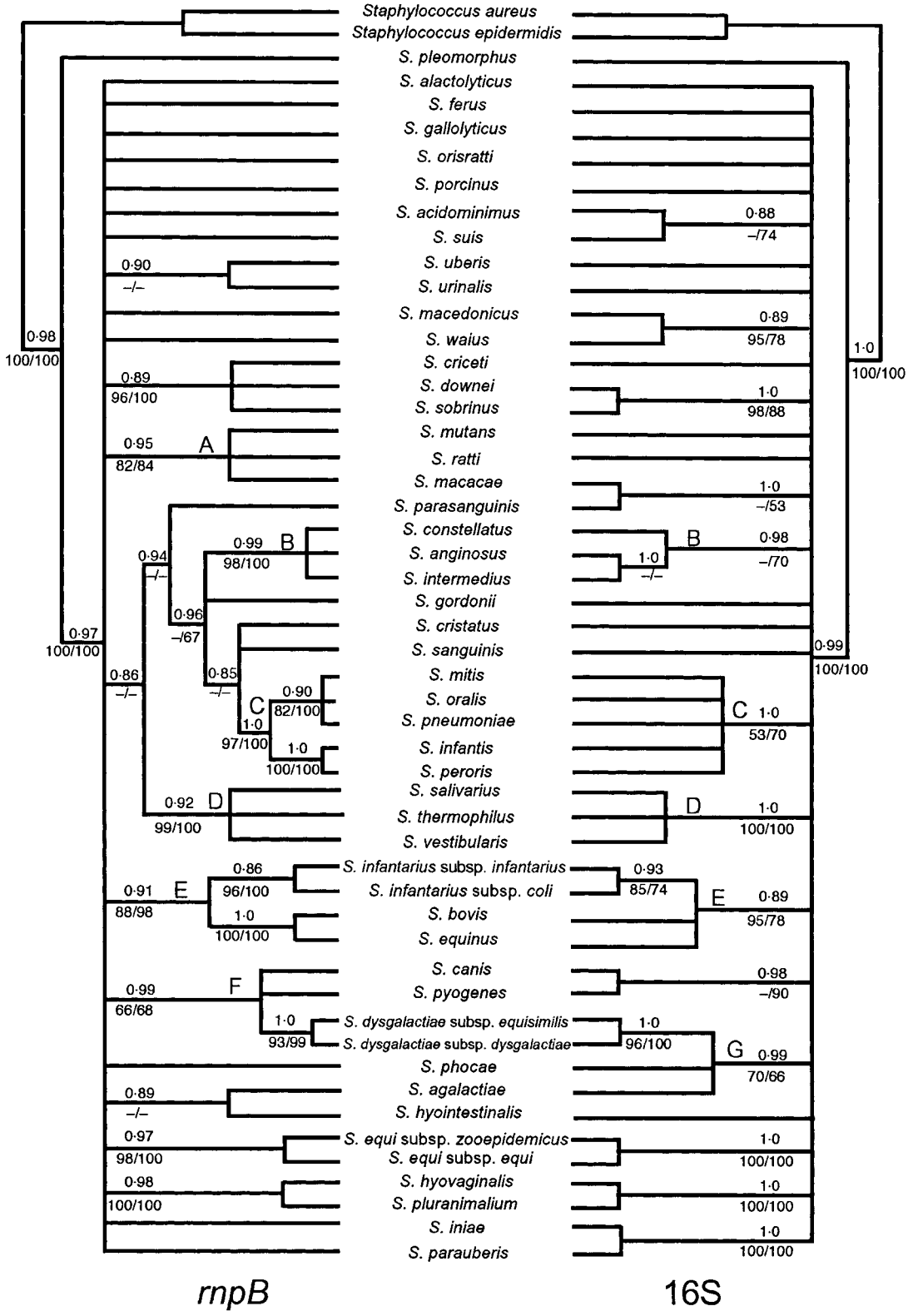

Fig. 2. Majority-rule consensus trees summarizing the results from Bayesian analysis of the *rnpB* (left) and 16S (right) datasets. Branches with a posterior probability of $<0{\cdot}85$ have been collapsed and hence relationships with a lower support will appear as polychotomies, even if they are present in the optimal tree. Numbers above branches indicate posterior probabilities of the branches in the Bayesian analysis; numbers below branches are bootstrap percentages for the MP and ME analyses, respectively. No value given (–) indicates a bootstrap proportion of $<50\,\%$, i.e. no support. Capital letters designate some clades discussed in the text.

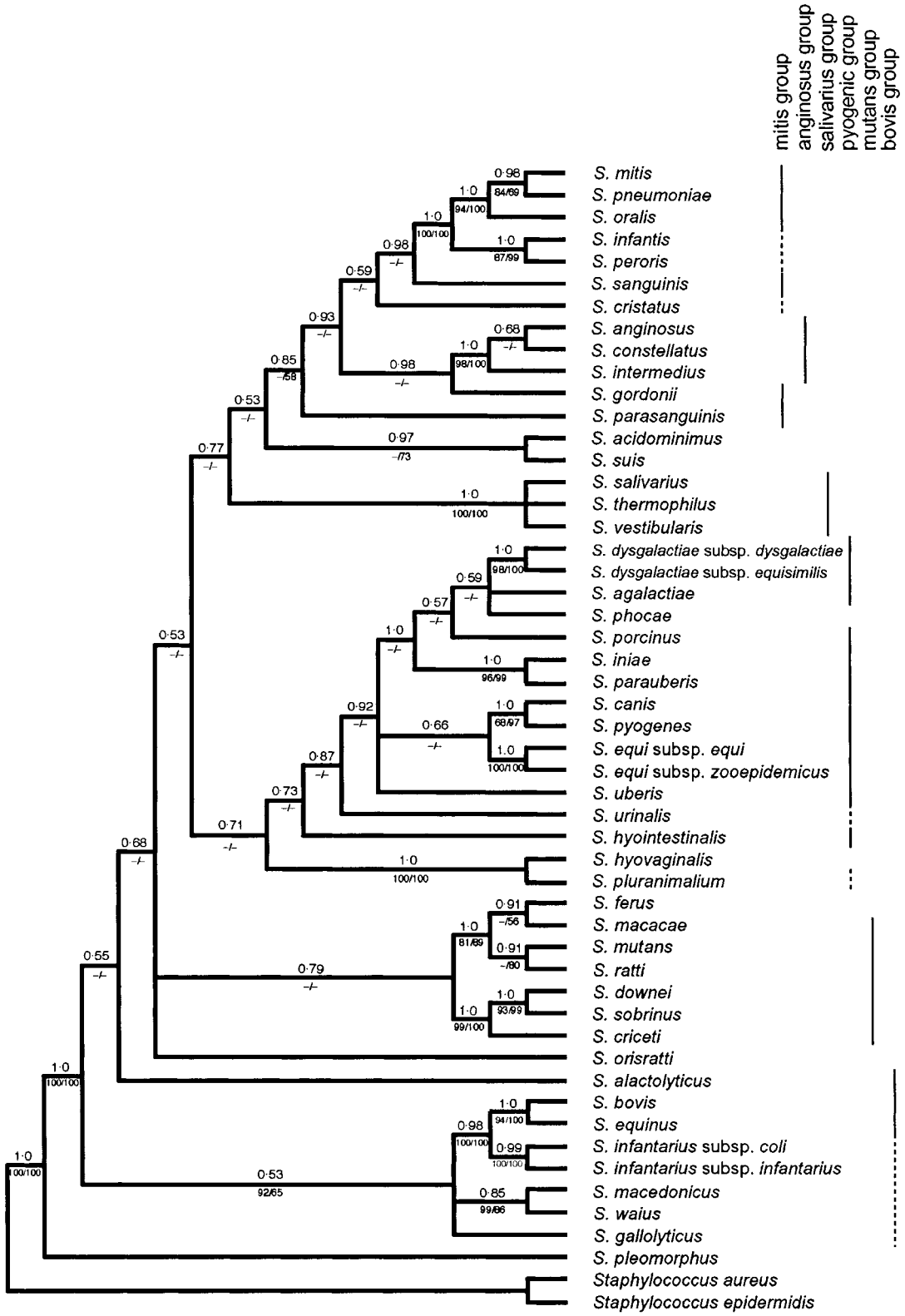

Fig. 3. Majority-rule consensus tree summarizing the result from Bayesian analysis of the combined (16S and *rnpB*) dataset. Branches with a posterior probability of $<0{\cdot}5$ have been collapsed. Numbers above branches are posterior probabilities; numbers below branches are bootstrap percentages for the MP and ME analyses, respectively. Bars indicate currently used taxonomic groups within *Streptococcus* (Kawamura *et al.*, 1995); solid bars indicate taxa present in the original classification, dashed bars indicate taxa added in subsequent analyses.

were distinguished (labelled B, C, D and G in Fig. 2), in addition to the clade that comprised all *Streptococcus* species except *S. pleomorphus*. The anginosus group (*Streptococcus anginosus*, *Streptococcus intermedius* and *Streptococcus constellatus*) and the salivarius group (*Streptococcus salivarius*, *Streptococcus thermophilus* and *Streptococcus vestibularis*) constitute small, specified clades (B and D, respectively), which were also supported in the analysis of *rnpB*. Interestingly, the anginosus clade was found to be a subclade in the same *rnpB* clade as members of the mitis group, as defined by Kawamura *et al.* (1999). Of the nine species in the mitis group, five (*Streptococcus mitis*, *S. oralis*, *Streptococcus pneumoniae*, *Streptococcus infantis* and *Streptococcus peroris*) constitute a well-supported clade in the 16S rRNA tree (clade C), whereas other species in this group either were not contained in a clade with substantial support (*Streptococcus gordonii*, *Streptococcus cristatus* and *Streptococcus sanguinis*) or, as is the case with *Streptococcus parasanguinis*, formed a clade with another species (*Streptococcus macacae*). The latter is one of the two cases of strong incongruence between the two genes; in the *rnpB* tree, *S. macacae* forms a well-supported clade with *Streptococcus mutans* and *Streptococcus ratti*. Low bootstrap support, however, for the *S. macacae*/*S. parasanguinis* clade in 16S analysis may indicate that the conflict is due to artefactual effects in the 16S analysis.

In the pyogenic group, there is support for *Streptococcus dysgalactiae*, *Streptococcus agalactiae* and *Streptococcus phocae* in one clade (G) in the 16S tree, although there is no support for *Streptococcus iniae* as the closest relative, as stated previously (Skaar *et al.*, 1994). The remaining eight species in this group according to the previous classification (Kawamura *et al.*, 1995) are either found in well-supported pair-clades (*Streptococcus canis* and *S. pyogenes*; *S. iniae* and *Streptococcus parauberis*) or without any well-supported association with other species (*Streptococcus hyointestinalis*, *Streptococcus porcinus*, *Streptococcus uberis* and *Streptococcus equi*). In contrast, in the *rnpB* tree, *S. dysgalactiae*, *S. canis* and *S. pyogenes* form a clade (F), whereas *S. agalactiae* do not form a well-supported clade with any other streptococcal species. This is the other strong incongruence between the 16S and *rnpB* datasets, for which additional data will be needed before the issue can be resolved. It is worth noting, however, that analysis of the *sodA* gene (Poyart *et al.*, 1998; Whatmore & Whiley, 2002) is congruent with *rnpB* in this matter. The close relationship between the two type strains of *S. equi* subsp. *equi* and *S. equi* subsp. *zooepidemicus* was supported strongly in both the *rnpB* and 16S rRNA trees. Other species that were, until recently, considered to belong to the pyogenic group (*Streptococcus urinalis* and *S. hyointestinalis*, and the pair *Streptococcus pluranimalium* and *Streptococcus hyovaginalis*) did not receive substantial support for inclusion in such a clade (although there was no contradiction).

For the bovis group, *rnpB* data supported a clade (E) with *S. bovis*, *Streptococcus equinus*, *Streptococcus infantarius* subsp. *infantarius* and *S. infantarius* subsp. *coli* (*Streptococcus lutetiensis*) together. This is also supported by recent analysis of the *sodA* gene (Poyart *et al.*, 2002), and, with slightly lower support ($p=0{\cdot}89$), by the 16S dataset.

The notable lack of support in the *rnpB* tree for a clade that consists of the two identical sequences of *S. macedonicus* and *S. waius* may seem paradoxical. *S. macedonicus* and *S. waius* do form a clade in the optimal phylogeny, regardless of the method used. Support, however, also depends on how distinct taxa in a clade are related to other taxa. In this case, just a few sites differ from other sequences in species that are closely related to *S. macedonicus*/*S. waius* in the analysis; the hypothesis that *S. macedonicus* and *S. waius* are close relatives due to sequence convergence cannot be ruled out completely (but there is even less support in favour of such a hypothesis).

Species that constitute the mutans group in the 16S analysis of Kawamura *et al.* (1995) (*S. mutans*, *S. ratti*, *S. macacae*, *Streptococcus downei*, *Streptococcus sobrinus* and *Streptococcus criceti*) obtained no significant support to form a clade in the 16S tree. However, in the *rnpB* tree, *S. mutans*, *S. ratti* and *S. macacae* form a well-supported clade. Another clade is formed by *S. downei*, *S. sobrinus* and *S. criceti*, with bootstrap values of 96 and 100 % in MP and ME analyses, respectively, albeit with a posterior probability of only 0·89.

Our phylogenetic analysis of streptococcal 16S genes indicates that for several species there is no strongly supported evolutionary relationship. As the rate of substitution in this gene is slow, this is not surprising. Although certain criteria for using 16S sequences have been adopted in many studies (<97 % sequence similarity enables the 16S rRNA gene to differentiate species; Stackebrandt & Goebel, 1994), taxonomic classification analysis must be based on more than a single gene and, in addition to genetic data, ecological data must be considered. Analysis of separate genes showed nine nodes with at least three terminal branches and significant posterior probabilities of $\geqslant 0{\cdot}95$; three unique nodes for *rnpB*, two unique nodes for 16S and four nodes common to both genes (Fig. 2). By combining the two genes in one analysis, the number of nodes increased to 12 (Fig. 3), indicating a higher resolution and increased clade support in phylogenetic analysis when available data are combined.

The taxonomically most significant result from the combined analysis (Fig. 3) is that the mitis group, as currently circumscribed, is not monophyletic without inclusion of the anginosus group. The latter formed a subclade within a clade that also comprised the mitis group, which is in contrast to analysis of the *sodA* gene (Whatmore & Whiley, 2002). Interestingly, all species in this clade (anginosus and mitis groups) have humans as their host organism. Combined analysis also showed weak support for the placement of *S. urinalis*, *S. hyointestinalis*, *S. hyovaginalis* and *S. pluranimalium* in the pyogenic clade. Furthermore, combined analysis firmly (posterior probability, 1·0) placed

Streptococcus ferus in the mutans group (with *S. mutans*, *S. ratti* and *S. macacae*), in which it has been included until recently. Whatmore & Whiley (2002) found no support in the 16S rRNA or *sodA* genes for inclusion of *S. ferus* in the mutans (or any other established) group and concluded that it is distantly related to all other *Streptococcus* species.

To characterize the variability in the 16S rRNA and *rnpB* genes, we calculated the Shannon–Wiener information index for each site over the same set of taxa. The frequency of positions with low nucleotide variation is much higher for the 16S rRNA gene than for the *rnpB* gene. On average, the index is three times higher for *rnpB* (0·45) than for 16S rRNA (0·15) (Fig. 4). A frequently used estimate of the variation in genes is obtained by comparing mean sequence similarity. However, this rough estimate does not take into account the fact that genes may not differentiate species properly if the nucleotide variation is limited to relatively few positions. Nevertheless, our analysis shows that for streptococci, *rnpB* has higher potential for species discrimination than the 16S rRNA gene. In fact, separate bacterial species have been found to have identical 16S rRNA gene sequences (Fox *et al.*, 1992). Another advantage of *rnpB* is the single-copy expression of the gene. In contrast, the 16S rRNA gene may have several copy variants in the genome and sequence heterogeneities may result in erroneous genotyping (Nubel *et al.*, 1996).

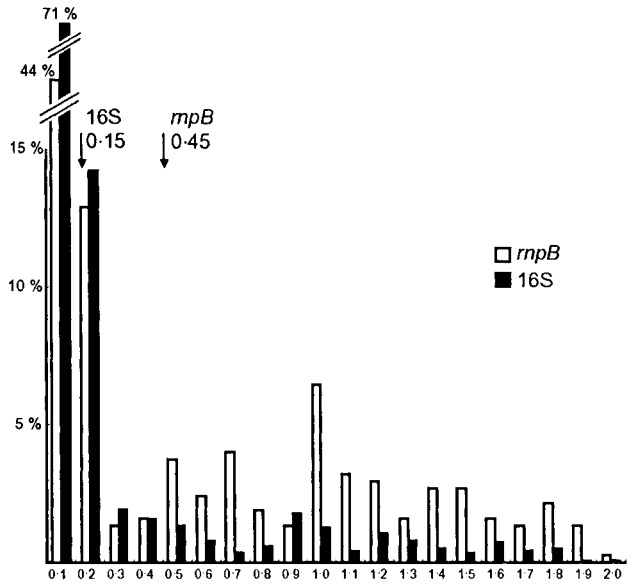

Fig. 4. Analysis of nucleotide variation of the *rnpB* and 16S genes in 50 type strains of streptococcal species. The Shannon–Wiener (S–W) information index expresses variation as information at each nucleotide position. Percentage of positions with different index values are plotted on the x-axis, with mean values for *rnpB* and 16S above. A site with the same base in all sequences has a S–W index of 0, whereas a site with all bases having the same frequency (0·25) obtains the maximum S–W value of 2.

Assessment of intraspecies proximity was performed on *rnpB* sequences from 15 strains of the three species in the salivarius group and from 18 strains of the three species in the anginosus group. Principal coordinate analysis showed that three distinct clusters, corresponding to nominal species, were formed for the salivarius group (Fig. 5) and five strains of *S. constellatus* formed one cluster. The two strains that represent the type strain of *S. intermedius* (i.e. equivalent strains with different histories) were identical, but were well-separated from the other

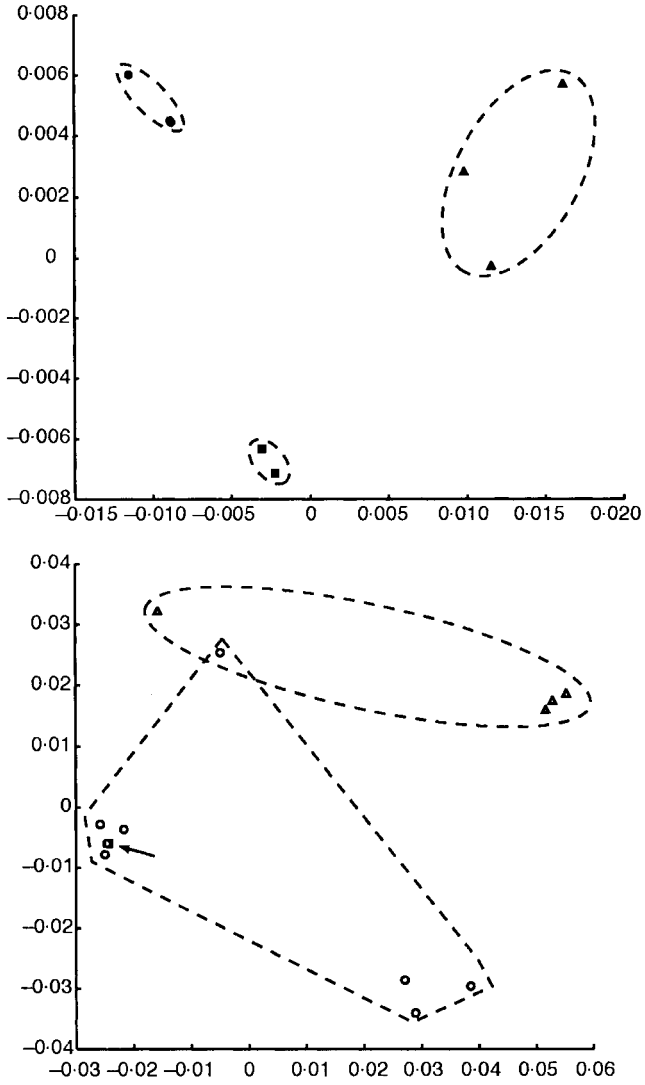

Fig. 5. First two principal coordinates resulting from principal coordinate analysis of *rnpB* sequences of (top) five strains of each species in the salivarius group [*S. salivarius* (▲, $n=5$), *S. thermophilus* (●, $n=5$) and *S. vestibularis* (■, $n=5$)] and (bottom) 18 strains of species in the anginosus group [*S. anginosus* (○, $n=8$), *S. constellatus* (□, $n=5$) and *S. intermedius* (△, $n=8$)]. Relative difference in sequence similarity between strains is shown for the 33 strains. Arrow indicates all *S. constellatus* strains and one *S. anginosus* strain.

three strains of this species. This is in agreement with recent findings of at least two populations of *S. intermedius* strains (Jacobs *et al.*, 2000a). Strains of *S. anginosus* formed two clusters and one additional strain was separated from the other strains. This heterogeneity has been reported previously and it has been suggested the type strain should be replaced (Whiley *et al.*, 1997; Bartie *et al.*, 2000; Jacobs *et al.*, 2000b). Thus, when intraspecies heterogeneity is seen in the *rnpB* sequence, it may reflect incomplete taxonomic classification.

In summary, putative secondary structures of RNase P RNA in 50 streptococcal species agreed with previously suggested models of interaction, although a few alterations were noted. The *rnpB* gene was shown to be suitable for phylogenetic analysis of closely related taxa in this study. Compared to the 16S rRNA gene, *rnpB* has a higher information value per nucleotide position that, in combination with its short length, gives it potential for species discrimination.

ACKNOWLEDGEMENTS

The *rnpB* sequences presented here were generated by Bioimics AB, Uppsala, Sweden. We thank Leif Kirsebom for reading the manuscript.

REFERENCES

Adnan, S., Li, N., Miura, H., Hashimoto, Y., Yamamoto, H. & Ezaki, T. (1993). Covalently immobilized DNA plate for luminometric DNA-DNA hybridization to identify viridans streptococci in under 2 hours. *FEMS Microbiol Lett* **106**, 139–142.

Altman, S. & Kirsebom, L. A. (1999). Ribonuclease P. In *The RNA World*, pp. 351–380. Edited by R. F. Gesteland, T. R. Cech and J. F. Atkins. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Baele, M., Storms, V., Haesebrouck, F., Devriese, L. A., Gillis, M., Verschraegen, G., de Baere, T. & Vaneechoutte, M. (2001). Application and evaluation of the interlaboratory reproducibility of tRNA intergenic length polymorphism analysis (tDNA-PCR) for identification of *Streptococcus* species. *J Clin Microbiol* **39**, 1436–1442.

Bartie, K. L., Wilson, M. J., Williams, D. W. & Lewis, M. A. O. (2000). Macrorestriction fingerprinting of ''*Streptococcus milleri*'' group bacteria by pulsed-field gel electrophoresis. *J Clin Microbiol* **38**, 2141–2149.

Beighton, D., Hardie, J. M. & Whiley, R. A. (1991). A scheme for the identification of viridans streptococci. *J Med Microbiol* **35**, 367–372.

Bentley, R. W. & Leigh, J. A. (1995). Development of PCR-based hybridization protocol for identification of streptococcal species. *J Clin Microbiol* **33**, 1296–1301.

Bentley, R. W., Leigh, J. A. & Collins, M. D. (1991). Intrageneric structure of *Streptococcus* based on comparative analysis of small-subunit rRNA sequences. *Int J Syst Bacteriol* **41**, 487–494.

Brown, J. W. (1999). The Ribonuclease P Database. *Nucleic Acids Res* **27**, 314.

De Gheldre, Y., Vandamme, P., Goossens, H. & Struelens, M. J. (1999). Identification of clinically relevant viridans streptococci and analysis of transfer DNA intergenic spacer length polymorphism. *Int J Syst Bacteriol* **49**, 1591–1598.

Farrow, J. A. E. & Collins, M. D. (1984). Taxonomic studies on streptococci of serological groups C, G and L and possibly related taxa. *Syst Appl Microbiol* **5**, 483–493.

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.

Garnier, F., Gerbaud, G., Courvalin, P. & Galimand, M. (1997). Identification of clinically relevant viridans group streptococci to the species level by PCR. *J Clin Microbiol* **35**, 2337–2341.

Gillespie, B. E., Jayarao, B. M. & Oliver, S. P. (1997). Identification of *Streptococcus* species by randomly amplified polymorphic deoxyribonucleic acid fingerprinting. *J Dairy Sci* **80**, 471–476.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338.

Gu, X., Fu, Y.-X. & Li, W.-H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol* **12**, 546–557.

Haas, E. S. & Brown, J. W. (1998). Evolutionary variation in bacterial RNase P RNAs. *Nucleic Acids Res* **26**, 4093–4099.

Haas, E. S., Banta, A. B., Harris, J. K., Pace, N. R. & Brown, J. W. (1996). Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. *Nucleic Acids Res* **24**, 4775–4782.

Herrmann, B., Winqvist, O., Mattsson, J. G. & Kirsebom, L. A. (1996). Differentiation of *Chlamydia* spp. by sequence determination and restriction endonuclease cleavage of RNase P RNA genes. *J Clin Microbiol* **34**, 1897–1902.

Herrmann, B., Pettersson, B., Everett, K. D. E., Mikkelsen, N. E. & Kirsebom, L. A. (2000). Characterization of the *rnpB* gene and RNase P RNA in the order *Chlamydiales*. *Int J Syst Evol Microbiol* **50**, 149–158.

Hillmann, J. D., Andrews, S. W., Painter, S. & Stashenko, P. (1989). Adaptive changes in a strain of *Streptococcus mutans* during colonization of the human oral cavity. *Microb Ecol Health Dis* **2**, 231–239.

Huelsenbeck, J. P. & Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu Rev Ecol Syst* **28**, 437–466.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.

Jacobs, J. A., Schot, C. S., Bunschoten, A. E. & Schouls, L. M. (1996). Rapid species identification of ''*Streptococcus milleri*'' strains by line blot hybridization: identification of a distinct 16S rRNA population closely related to *Streptococcus constellatus*. *J Clin Microbiol* **34**, 1717–1721.

Jacobs, J. A., Schot, C. S. & Schouls, L. M. (2000a). Haemolytic activity of the '*Streptococcus milleri* group' and relationship between haemolysis restricted to human red blood cells and pathogenicity in *S. intermedius*. *J Med Microbiol* **49**, 55–62.

Jacobs, J. A., Schot, C. S. & Schouls, L. M. (2000b). The *Streptococcus anginosus* species comprises five 16S rRNA ribogroups with different phenotypic characteristics and clinical relevance. *Int J Syst Evol Microbiol* **50**, 1073–1079.

Jayarao, B. M., Dore, J. J., Jr & Oliver, S. P. (1992). Restriction fragment length polymorphism analysis of 16S ribosomal DNA of *Streptococcus* and *Enterococcus* species of bovine origin. *J Clin Microbiol* **30**, 2235–2240.

Kawamura, Y., Hou, X.-G., Sultana, F., Miura, H. & Ezaki, T. (1995). Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. *Int J Syst Bacteriol* **45**, 406–408.

Kawamura, Y., Whiley, R. A., Shu, S.-E., Ezaki, T. & Hardie, J. M. (1999). Genetic approaches to the identification of the mitis group within the genus *Streptococcus*. *Microbiology* **145**, 2605–2613.

Kikuchi, K., Enari, T., Totsuka, K. & Shimizu, K. (1995). Comparison of phenotypic characteristics, DNA-DNA hybridization results, and results with a commercial rapid biochemical and enzymatic reaction system for identification of viridans group streptococci. *J Clin Microbiol* **33**, 1215–1222.

Kilian, M., Mikkelsen, L. & Henrichsen, J. (1989). Taxonomic studies of viridans streptococci: description of *Streptococcus gordonii* sp. nov. and emended descriptions of *Streptococcus sanguis* (White and Niven 1946), *Streptococcus oralis* (Bridge and Sneath 1982), and *Streptococcus mitis* (Andrewes and Horder 1906). *Int J Syst Bacteriol* **39**, 471–484.

Lawrence, J., Yajko, D. M. & Hadley, W. K. (1985). Incidence and characterization of beta-hemolytic *Streptococcus milleri* and differentiation from *S. pyogenes* (group A), *S. equisimilis* (group C), and large-colony group G streptococci. *J Clin Microbiol* **22**, 772–777.

Legendre, P. & Anderson, M. J. (1998). DISTPCOA program. Département de Sciences Biologiques, Université de Montréal.

Legendre, P. & Anderson, M. J. (1999). Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. *Ecol Monogr* **69**, 1–24.

Ludwig, W., Weizenegger, M., Kilpper-Bälz, R. & Schleifer, K. H. (1988). Phylogenetic relationships of anaerobic streptococci. *Int J Syst Bacteriol* **38**, 15–18.

Maidak, B. L., Cole, J. R., Lilburn, T. G. & 7 other authors (2001). The RDP-II (Ribosomal Database Project). *Nucleic Acids Res* **29**, 173–174.

Manachini, P. L., Flint, S. H., Ward, L. J. H., Kelly, W., Fortina, M. G., Parini, C. & Mora, D. (2002). Comparison between *Streptococcus macedonicus* and *Streptococcus waius* strains and reclassification of *Streptococcus waius* (Flint *et al.* 1999) as *Streptococcus macedonicus* (Tsakalidou *et al.* 1998). *Int J Syst Evol Microbiol* **52**, 945–951.

Massire, C., Jaeger, L. & Westhof, E. (1998). Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *J Mol Biol* **279**, 773–793.

Nubel, U., Engelen, B., Felske, A., Snaidr, J., Wieshuber, A., Amann, R. I., Ludwig, W. & Backhaus, H. (1996). Sequence heterogeneities of genes encoding 16S rRNAs in *Paenibacillus polymyxa* detected by temperature gradient gel electrophoresis. *J Bacteriol* **178**, 5636–5643.

Pannucci, J. A., Haas, E. S., Hall, T. A., Harris, J. K. & Brown, J. W. (1999). RNase P RNAs from some Archaea are catalytically active. *Proc Natl Acad Sci U S A* **96**, 7803–7808.

Posada, D. & Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818.

Poyart, C., Quesne, G., Coulon, S., Berche, P. & Trieu-Cuot, P. (1998). Identification of streptococci to species level by sequencing the gene encoding the manganese-dependent superoxide dismutase. *J Clin Microbiol* **36**, 41–47.

Poyart, C., Quesne, G. & Trieu-Cuot, P. (2002). Taxonomic dissection of the *Streptococcus bovis* group by analysis of manganese-dependent superoxide dismutase gene (*sodA*) sequences: reclassification of '*Streptococcus infantarius* subsp. *coli*' as *Streptococcus lutetiensis* sp. nov. and of *Streptococcus bovis* biotype II.2 as *Streptococcus pasteurianus* sp. nov. *Int J Syst Evol Microbiol* **52**, 1247–1255.

Rodriguez, F., Oliver, J. L., Marin, A. & Medina, J. R. (1990). The general stochastic model of nucleotide substitution. *J Theor Biol* **142**, 485–501.

Rudney, J. D. & Larson, C. J. (1994). Use of restriction fragment polymorphism analysis of rRNA genes to assign species to unknown clinical isolates of oral viridans streptococci. *J Clin Microbiol* **32**, 437–443.

Saruta, K., Matsunaga, T., Hoshina, S., Kono, M., Kitahara, S., Kanemoto, S., Sakai, O. & Machida, K. (1995). Rapid identification of *Streptococcus pneumoniae* by PCR amplification of ribosomal DNA spacer region. *FEMS Microbiol Lett* **132**, 165–170.

Schmidhuber, S., Ludwig, W. & Schleifer, K. H. (1988). Construction of a DNA probe for the specific identification of *Streptococcus oralis*. *J Clin Microbiol* **26**, 1042–1044.

Shannon, C. E. & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press.

Skaar, I., Gaustad, P., Tønjum, T., Holm, B. & Stenwig, H. (1994). *Streptococcus phocae* sp. nov., a new species isolated from clinical specimens from seals. *Int J Syst Bacteriol* **44**, 646–650.

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.

Swofford, D. L. (2000). PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), release 4 for Apple Macintosh, Intel, Linux and SGI/Irix. Sunderland, MA: Sinauer Associates.

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**, 512–526.

Tanner, M. A. & Cech, T. R. (1995). An important RNA tertiary interaction of group I and group II introns is implicated in gram-positive RNase P RNAs. *RNA* **1**, 349–350.

Tardif, G., Sulavik, M. C., Jones, G. W. & Clewell, D. B. (1989). Spontaneous switching of the sucrose-promoted colony phenotype in *Streptococcus sanguis*. *Infect Immun* **57**, 3945–3948.

Waddell, P. J. & Penny, D. (1996). Evolutionary trees of apes and humans from DNA sequences. In *Handbook of Human Symbolic Evolution*, pp. 53–73. Edited by A. Lock & C. R. Peters. Oxford: Oxford University Press.

Whatmore, A. M. & Whiley, R. A. (2002). Re-evaluation of the taxonomic position of *Streptococcus ferus*. *Int J Syst Evol Microbiol* **52**, 1783–1787.

Whiley, R. A., Duke, B., Hardie, J. M. & Hall, L. M. C. (1995). Heterogeneity among 16S–23S rRNA intergenic spacers of species within the '*Streptococcus milleri* group'. *Microbiology* **141**, 1461–1467.

Whiley, R. A., Hall, L. M. C., Hardie, J. M. & Beighton, D. (1997). Genotypic and phenotypic diversity within *Streptococcus anginosus*. *Int J Syst Bacteriol* **47**, 645–650.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: with Engineering Applications*. New York: Wiley.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol* **10**, 1396–1401.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**, 306–314.