Partial sequence comparison of the rpoB, sodA, groEL and gyrB genes within the genus Streptococcus

Olga O. Glazunova, Didier Raoult and Véronique Roux

Laboratoire de Bactériologie – Virologie, Hôpital de la Timone, CNRS UMR 6236, IFR48, 264 rue Saint-Pierre, 13385 Marseille, Cedex 05, France

Phylogenetic analysis and species identification of members of the genus Streptococcus were carried out using partial sequence comparison of the 16S rRNA gene (1468–1478 bp), rpoB, encoding the β subunit of RNA polymerase (659–680 bp), sodA, encoding the manganese-dependent superoxide dismutase (435–462 bp), groEL, encoding the 60 kDa heat-shock protein (757 bp), and gyrB, encoding the B subunit of DNA gyrase (458–461 bp). For the first time, most species within the genus Streptococcus were represented in the study (65 strains, representing 58 species and nine subspecies). Phylogenies inferred from rpoB, sodA, gyrB and groEL sequence comparisons were more discriminative than those inferred from 16S rRNA gene sequence comparison, and showed common clusters. The minimal interspecies divergence was 0.3, 2.7, 0, 2.5 and 3.4 % for the 16S rRNA gene, rpoB, sodA, gyrB and groEL, respectively. In general, groEL partial gene sequence comparison represented the best tool for identifying species and subspecies and for phylogenetic analysis.

The streptococci are non-motile, Gram-positive cocci that are a part of the family Streptococcaceae. Currently, 66 species and 12 subspecies are recognized as members of the genus Streptococcus (http://www.bacterio.cict.fr/s/streptococcus.html). Until recently, streptococci were identified based only on the haemolytic reaction, group carbohydrate antigens and phenotypic tests (Facklam, 2002). Phylogenetic analysis and identification of representatives of the genus Streptococcus has been based on 16S rRNA gene sequence comparisons (Bentley et al., 1991), but the lack of variability of the 16S rRNA gene sequence did not allow identification of closely related species or subspecies. Other genes have subsequently been used (Garnier et al., 1997; Kawamura et al., 1999, 2005; Igarashi et al., 2001; Teng et al., 2002; Tapp et al., 2003; Picard et al., 2004; Hung et al., 2005).

In order to discriminate among species, a gene must be present in a single copy, widely distributed among bacterial genomes and sufficiently variable to differentiate species of the considered genus (Zeigler, 2003). We chose to complete the work initiated by Poyart et al. (1998), Drancourt et al. (2004) and Hoshino et al. (2005), and determined 16S rRNA gene sequences for streptococcal type strains that were not represented in GenBank and sequenced gyrB and groEL gene fragments. During the course of our study, Itoh et al. (2006) published gyrB gene sequences for 37 species. In our study, most recognized species of the genus Streptococcus were included, for the first time. We compared the sequence data for each gene for phylogenetic analysis and to evaluate the discriminatory power of the individual genes in identification of streptococci.

We studied 65 streptococcal strains (representing 58 species and nine subspecies; see Supplementary Table S1, available in IJSEM Online), mostly obtained from the Pasteur Institute Collection (CIP, Paris, France). Bacterial DNA was extracted using the MagNA Pure LC DNA isolation kit III (Roche) with the MagNA Pure LC instrument as described by the manufacturer. PCR amplification of the 16S rRNA gene was performed using the universal primer pair fD1 and rp2 (Weisburg et al., 1991). The primer pair d1 and d2 (Poyart et al., 1998) and primers StreptoF and StreptoR (Drancourt et al., 2004) were respectively used for PCR amplification of sodA and rpoB, as described previously. The locations of the amplification primers for groEL and gyrB genes (streptogroELd/streptogroELr and streptogyrBd/streptogyrBr) were chosen in conserved regions. Primers are described in Supplementary Table S2. At the outset of this study, we defined these conserved regions by aligning all the available genome sequences for streptococci [Streptococcus pneumoniae] strains R6 (GenBank accession
no. NC_003098) and TIGR4 (NC_003028), *Streptococcus agalactiae* strains NEM316 (NC_004368) and 2603V/R (NC_004116), *Streptococcus pyogenes* strains M1 GAS (NC_002737) and MGAS315 (NC_004070), *Streptococcus mutans* UA159 (NC_004350) and *Streptococcus thermophilus* strains LMG 18311 (NC_006448) and CNRZ1066 (NC_006449)] and other phylogenetically related bacteria [*Lactococcus lactis* subsp. *lactis* II1403 (NC_002662), *Listeria innocua* Clip11262 (NC_003212) and *Listeria monocytogenes* EGD-e (NC_003210)]. For *groEL*, the sequences determined by Teng *et al.* (2002) and Hung *et al.* (2005) were also included in the alignment. The annealing temperature for the PCR was 52 °C. No product was obtained for *gyrB* with the chosen primer pair for the type strain of *Streptococcus suis*; another pair of primers was designed in a conserved downstream region (GyrBsuisd and GyrBsuisr), and a PCR product was obtained using StreptogyrBd and GyrBsuir. Other primers were designed for sequencing reactions. PCR products were purified using MultiScreen PCR plates (Millipore) and sequenced using a DNA sequencing kit (BigDye Terminator v. 1.1 cycle sequencing kit; PE Biosystems) according to the manufacturer’s instructions. Sequence products were purified and electrophoresis was performed with an Applied Biosystems 3100 Genetic Analyzer. The sequences were submitted to GenBank, and accession numbers are indicated in Supplementary Table S1. Gene sequences were aligned using the multi-sequence alignment program CLUSTAL_X 1.8 (Thompson *et al.*, 1997). Phylogenetic relationships were determined by using the program MEGA version 4 (Tamura *et al.*, 2007). Distance matrices were determined following the assumptions described by Kimura (1980) and were used to elaborate dendrograms using the neighbour-joining method (Saitou & Nei, 1987). The maximum-parsimony algorithm (Farris, 1970) and the maximum-likelihood method using the program DNAML in the PHYLIP software package (Felsenstein, 1989) were also used for phylogenetic analysis. Bootstrap values were obtained for a consensus tree based on 100 randomly generated trees.

The genes studied represented different COG functional categories: replication, recombination and repair genes (*gyrB; COG0187*), transcription genes (*rpoB; COG0085*), post-translational modification, protein turnover and chaperone genes (*groEL; COG0459*) and inorganic ion transport and metabolism genes (*sodA; COG0605*). Moreover, three of the four chosen genetic targets have been considered as part of the ‘minimal gene-set’ (*i.e.* genes necessary to maintain a living cell; Gil *et al.*, 2004). These are distributed in different categories: DNA metabolism (*gyrB*), RNA metabolism (*rpoB*) and protein processing, folding and secretion (*groEL*).

Phylogenies were inferred from the five analysed genes and from the 16S rRNA gene. Identical topologies were obtained for each gene using the neighbour-joining, maximum-parsimony and maximum-likelihood methods; however, the highest bootstrap values were obtained with the neighbour-joining method, and it is these values that are referred to in the discussion. We considered only clusters supported by bootstrap values greater than 85 %.

A PCR fragment of 1468–1478 bp was amplified from the 16S rRNA gene. In the ‘mitis’ group, five species grouped together with a bootstrap value of 91 % (*S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis*, *S. oralis* and *S. infantis*) (Supplementary Fig. S1). *Streptococcus oligofermentans* and *S. sinensis* clustered together with a significant bootstrap value of 99 %. Seven other clusters were noted: (i) the ‘salivarius’ group (*Streptococcus thermophilus*, *S. vestibularis* and *S. salivarius*); (ii) representatives of the ‘equinus’ group (*Streptococcus galalacticus* subsp. *galalcyticus*, *S. galalcyticus* subsp. *pasteurianus*, *S. galalcyticus* subsp. *macedonicus*, *S. infantarius* subsp. *infantarius*, *S. infantarius* subsp. *coli* and *S. equinus*); (iii) *Streptococcus downei* and *S. sobrinus*; (iv) *S. pyogenes* and *S. canis*; (v) *Streptococcus intermedius* and *S. constellatus*; (vi) *Streptococcus parauberis* and *S. iniae*; and (vii) the ‘hyovaginalis’ group (*Streptococcus hyovaginalis*, *S. pluranimalium* and *S. thoralensis*). All the other species clustered with non-significant bootstrap values at the nodes.

A fragment of 435 bp was obtained for all the streptococci except *Streptococcus massiliensis* (462 bp) for *sodA*. The PCR fragment represented 69.5 % of the *sodA* gene sequence. All the representatives of the ‘mitis’ group clustered with a bootstrap value of 97 % (Supplementary Fig. S2a). Three subclusters were found: the first included *Streptococcus anginosus*, *S. intermedius* and *S. constellatus* (bootstrap value 99 %), the second included *Streptococcus sanguinis* and *S. gordonii* (bootstrap value 87 %) and the third included *S. infantis*, *S. australis*, *S. peroris*, *S. oralis*, *S. pneumoniae*, *S. pseudopneumoniae*, *S. mitis* and *S. oligofermentans* (bootstrap value 99 %). Six other clusters were noted: (i) the ‘salivarius’ group; (ii) the ‘equinus’ group; (iii) the ‘hyovaginalis’ group; (iv) the ‘criciet’ group (*Streptococcus criciet*, *S. downei* and *S. sobrinus*); (v) the ‘pyogenes’ group (*S. canis*, *S. pyogenes*, *S. dysgalactiae* subsp. *dysgalactiae* and *S. dysgalactiae* subsp. *equisimilis*); and (vi) *Streptococcus devriesii* and *S. ratti*. All the other species clustered with non-significant bootstrap values at the nodes.

A fragment of 680 bp was amplified and sequenced for *rpoB* for all species except *Streptococcus marinimammalium* (659 bp). The PCR fragment represented 18.2 % of the full-length *rpoB* gene sequence. In the ‘mitis’ group, seven species were grouped together with a significant bootstrap value of 100 % (*S. infantis*, *S. peroris*, *S. oralis*, *S. oligofermentans*, *S. mitis*, *S. pneumoniae* and *S. pseudopneumoniae*) (Supplementary Fig. S2b). *Streptococcus constellatus*, *S. anginosus* and *S. intermedius* formed one monophyletic cluster and *S. sinensis*, *S. sanguinis* and *S. gordonii* formed another, with respective bootstrap values of 100 and 98 %. *S. australis* and *S. parasanguinis* clustered together with a bootstrap value of 100 %. Five other
clusters were noted: (i) the ‘salivarius’ group; (ii) the ‘equinus’ group; (iii) the ‘hyovaginalis’ group; (iv) the ‘criceti’ group; and (v) the ‘pyogenes’ group. Three pairs of species grouped together with significant bootstrap values: (i) *Streptococcus ovis* and *S. minor*; (ii) *S. devriesei* and *S. ratti*; and (iii) *S. suis* and *S. gallinarum*. All the other species clustered with non-significant bootstrap values at the nodes.

Primers used to amplify *gyrB* generated a 458 bp fragment (representing 23.5% of the whole gene sequence) for all species except *Streptococcus fœrus* (461 bp). The representatives of the ‘mitis’ group were separated into several clusters (Supplementary Fig. S2c). The first included *S. constellatus*, *S. intermedius* and *S. anginosus* (bootstrap value 100%). The second included *S. massiliensis*, *S. sinensis*, *S. sanguinis* and *S. cristatus* (bootstrap value 90%) and the third included *S. infantis*, *S. peroris*, *S. oralis*, *S. oligofermentans*, *S. mitis*, *S. pneumoniae* and *S. pseudopneumoniae* (bootstrap value 100%). Another cluster including *S. gallinarum*, *S. ovis*, *S. minor*, *S. acidominimus* and *S. suis* (bootstrap value 42%) split the ‘mitis’ group streptococci. Using the maximum-parsimony method, *S. australis* clustered with the ‘minor’ subcluster, but, using the neighbour-joining and maximum-likelihood methods, *S. australis* clustered with the ‘sanguinis’ subcluster. Five other clusters were noted: (i) the ‘salivarius’ group; (ii) the ‘hyovaginalis’ group; (iii) the ‘criceti’ group; (iv) the ‘pyogenes’ group and *Streptococcus castoreus*; and (v) *S. devriesei*, *S. ratti* and *S. mutans*. For the representatives of the ‘equinus’ group, two subgroups were noted: the first included the subspecies of *S. galolyticus* and the second included *S. equinus* and the subspecies of *S. infantarius*. *Streptococcus alactolyticus* was closely related to the latter species, but the bootstrap value for the node was not significant (70%). All the other species clustered with non-significant bootstrap values at the nodes.

Previously, Itoh et al. (2006) performed a phylogenetic analysis of 37 species of the genus *Streptococcus* based on sequence comparison of a 900 bp *gyrB* gene fragment. Many clusters were the same in our study: the ‘salivarius’ group, the ‘criceti’ group, *S. mutans* and *S. ratti*, the ‘pyogenes’ group and *Streptococcus phocae*, and *S. hyovaginalis* and *S. thoraltensis*. *S. equinus* and *S. galolyticus* clustered together. All the species of the ‘mitis’ group clustered together with an organization similar to that found in our study, but more species were included in our study.

In all studied strains, a 757 bp fragment was amplified for *groEL*, representing 46.6% of the full-length sequence. All the representatives of the ‘mitis’ group clustered with a bootstrap value of 86% (Fig. 1). *S. acidominimus* and *S. suis* were also included in this cluster. Two subclusters were observed: the first included *S. constellatus*, *S. anginosus*, *S. intermedius*, *S. massiliensis*, *S. sanguinis*, *S. gordonii*, *S. cristatus* and *S. sinensis* (bootstrap value 89%) and the second included *S. australis*, *S. parasanguinis*, *S. peroris*, *S. infantis*, *S. oralis*, *S. oligofermentans*, *S. mitis*, *S. pneumoniae* and *S. pseudopneumoniae* (bootstrap value 94%). Six other clusters were noted: (i) the ‘salivarius’ group; (ii) the ‘equinus’ group; (iii) the ‘hyovaginalis’ group; (iv) the ‘criceti’ group; (v) the ‘pyogenes’ group and *S. castoreus*; and (vi) *S. devriesei*, *S. ratti*, *S. mutans* and *S. macacae*. All the other species clustered with non-significant bootstrap values at the nodes.

A phylogenetic analysis based on *groEL* gene sequence comparison was performed by Teng et al. (2002) for viridans group streptococci (*S. constellatus*, *S. anginosus*, *S. intermedius*, *S. oralis*, *S. pneumoniae*, *S. mitis*, *S. sanguinis*, *S. gordonii*, *S. mutans*, *S. salivarius* and *S. bovis*). Our results are in agreement with their conclusions. The three species of the ‘anginosus’ group clustered together. *S. mitis*, *S. oralis* and *S. pneumoniae* were highly related. *S. sanguinis* and *S. gordonii* clustered together. It is difficult to compare the other groupings because only a few species were included in the study of Teng et al. (2002). Hung et al. (2005) performed a phylogenetic analysis based on *groEL* gene sequence comparison of the ‘mutans’ group streptococci (*S. sobrinus*, *S. downei*, *S. criceti*, *S. ratti* and *S. mutans*). All species clustered together, but with a non-significant bootstrap value of 56%, and two subclusters were noted. The first included the ‘criceti’ group and the second included *S. mutans* and *S. ratti*.

*rpoB*, *sodA*, *groEL*, *gyrB* and 16S rRNA gene-based phylogenetic trees inferred by using the neighbour-joining method were compared. The ‘equinus’ group was found in all the phylogenetic analyses. This cluster included four species, *S. galolyticus*, *S. infantarius*, *S. equinus* and *S. alactolyticus*, and the bootstrap value found for the representative node was significant except in the phylogenies inferred from 16S rRNA and *gyrB* sequence analysis. *S. equinus* clustered with *S. infantarius* with significant bootstrap values of 91–100%. The ‘salivarius’ group and the ‘hyovaginalis’ group were always found (bootstrap values 99–100 and 92–100%, respectively). The ‘criceti’ group was always found with a significant bootstrap value (90–100%) except in the 16S rRNA gene sequence analysis (50%). The ‘pyogenes’ group was also found except in the phylogenetic analysis inferred from 16S rRNA gene sequence comparison. *S. constellatus*, *S. intermedius* and *S. anginosus* clustered together (bootstrap value 99–100%), generally in the ‘mitis’ cluster, except in the tree obtained by 16S rRNA gene sequence comparison. The organization of the other representatives of the ‘mitis’ group was variable.

Recently, *S. equinus* and *S. bovis* were included as the same species, *S. equinus* (Schlegel et al., 2003). The two strains clustered together in all the phylogenetic analyses except for those from *gyrB* gene sequence comparisons. 16S rRNA gene sequence similarities from type strains of different species ranged from 88.8% (*S. macacae/S. acidominimus*) to 99.7% (*S. thermophilus/S. salivarius*), *sodA* sequence similarities ranged from 63.6% (*S. massi-
liensis/S. dysgalactiae subsp. dysgalactiae) to 100 % (S. pneumoniae/S. pseudopneumoniae), rpoB sequence similarities ranged from 76.1 % (S. gordonii/S. constellatus) to 97.6 % (S. thermophilus/S. vestibularis), gyrB sequence similarities ranged from 64.6 % (S. phocae/S. downei) to 97.5 % (S. constellatus/S. intermedius) and groEL sequence similarities ranged from 72.6 % (S. sobrinus/S. agalactiae, S. entericus/S. iniae) to 96.6 % (S. thermophilus/S. vestibularis). Supplementary Fig. S3 shows the percentage similarity between each pair of strains for each gene; it shows that rpoB is the least variable gene (92.5 % of the values are between 80 and 90 %) and gyrB is the most variable (82.1 % of the values are between 70 and 80 % and 15 % are <70 %).
16S rRNA gene sequence similarities from two subspecies ranged from 97.6% (S. gallokyticus subsp. gallokyticus/S. gallokyticus subsp. macedonicus) to 99.9% (S. infantarius subsp. infantarius/S. infantarius subsp. coli), sodA sequence similarities ranged from 88.0% (S. dysgalactiae subsp. dysgalactiae/S. dysgalactiae subsp. equi) to 98.9% (S. equi subsp. equi/S. equi subsp. zooepidemicus), rpoB sequence similarities ranged from 97.9% (S. dysgalactiae subsp. dysgalactiae/S. dysgalactiae subsp. equi) to 99.6% (S. gallokyticus subsp. gallokyticus/S. gallokyticus subsp. macedonicus), gyrB sequence similarities ranged from 93.7% (S. infantarius subsp. infantarius/S. infantarius subsp. coli) to 98.5% (S. gallokyticus subsp. gallokyticus/S. gallokyticus subsp. macedonicus) and groEL sequence similarities ranged from 93.9% (S. dysgalactiae subsp. dysgalactiae/S. dysgalactiae subsp. equi) to 98.8% (S. gallokyticus subsp. gallokyticus/S. gallokyticus subsp. macedonicus). Based on subspecies type strain sequence comparisons, the similarity is lowest for sodA and gyrB. Moreover, in the phylogenies inferred from sodA and gyrB gene sequence comparisons, the two subspecies of S. infantarius did not cluster together. Also, in the phylogeny inferred from sodA gene sequence comparison, the two subspecies of S. dysgalactiae did not cluster together. In general, groEL is the best tool for differentiating subspecies: the percentage of similarity is discriminatory and clustering of subspecies is always obtained in the phylogenetic analysis.

Phylogenetic analysis and molecular identification based on only one gene is not recommended because of possible gene duplication, lateral gene transfer or gene loss, which can distort the results. Partial sequencing of the 16S rRNA gene is the first step in bacterial strain identification. It allows identification at the genus level and the recognition of novel species if the similarity between the new isolate and recognized species is ≤97% (Stackebrandt et al., 2002). Housekeeping genes are also suitable for identification at the species or subspecies level, as demonstrated for several bacterial phyla (Holmes et al., 2004; Martens et al., 2008; Naser et al., 2007). In this study, the use of sodA, rpoB, gyrB and groEL gene sequence comparisons allowed us to improve our knowledge of the taxonomy and phylogenetic relationships of the genus Streptococcus. Presently, sodA and rpoB gene sequence comparisons are the tools used to identify strains at the species level in the genus Streptococcus. gyrB has been shown to be a good tool for identifying some members of the genera Aeromonas (Küpfert et al., 2006), Salmonella, Shigella, Escherichia (Fukushima et al., 2002) and Pseudomonas (Radice et al., 2006). At present, the phylogeny obtained from gyrB gene sequence comparison presents some inconsistencies. groEL has also been described as a good tool for identification in the genera Rickettsia (Lee et al., 2003), Campylobacter (Kärenlampi et al., 2004), Lactobacillus (Blaiotta et al., 2008) and Bacillus (Chang et al., 2003). In our study, which included representatives of most species of the genus Streptococcus for the first time, groEL gene sequence comparison was a powerful tool for establishing phylogeny and differentiating Streptococcus species. However, for each species, sequences will have to be obtained from several isolates in order to confirm the results established by this study of type strains.

References


