Sequence of the gtfK gene of *Streptococcus salivarius* ATCC 25975 and evolution of the gtf genes of oral streptococci

PHILIP M. GIFFARD, DONNA M. ALLEN, CAROLYN P. MILWARD, CHRISTINE L. SIMPSON and NICHOLAS A. JACQUES*

Institute of Dental Research, United Dental Hospital, 2 Chalmers Street, Surry Hills, NSW 2010, Australia

(Received 7 September 1992; revised 7 December 1992; accepted 1 March 1993)

Many strains of oral streptococci secrete glucosyltransferases (GTFs) that polymerize sucrose into glucans that form an integral part of the plaque matrix on the tooth surface. Recently, we reported the cloning of two closely linked GTF-encoding genes (gtfJ and gtfK) from *Streptococcus salivarius* ATCC 25975 as well as the sequence of gtfJ, which encodes a primer-dependent GTF that synthesizes an insoluble product (a GTF-I). In this communication we report the sequence of gtfK, which encodes a primer-dependent GTF that synthesizes a soluble product (a GTF-S), as well as the sequence of a small downstream open reading frame of unknown function. The deduced sequence of GtfK was compared with those of seven other streptococcal Gtfs and an unrooted phylogenetic tree constructed. This analysis suggested that Gtfs with similar product specificities do not form phylogenetic clusters and was consistent with currently accepted phylogenetic schemes. The tree was tested by constructing a series of ‘sub-trees’ from different blocks of the alignment. Evidence was obtained for recombination events involving gtfB and gtfC from *S. mutans* GS-5, gtfJ and gtfK from *S. salivarius*, as well as the gtfI genes from *S. downei* and *S. sobrinus*. The recombination events between gtfB and gtfC, and between the two gtfI genes, were confirmed by examining divergences at silent sites.

Introduction

Members of the mutans and sanguis groups of streptococci, as well as *Streptococcus salivarius* strains, synthesize extracellular glucosyltransferases (GTFs) that are considered to be important virulence factors in the aetiology of dental caries. The GTFs are highly hydrophilic monomeric proteins of approximate $M_r$ 160000 that form glucans by the polymerization of the glucose moiety of sucrose and the release of fructose. It is these glucans that promote dental caries by conferring adherent and anti-diffusion properties on dental plaque (Gibbons, 1968). In most cases the GTFs require a primer glucan for activity. These primer-dependent GTFs either extend the length of, or add branching points to, a pre-existing glucan. It is the nature of the linkages in these glucans that determines their solubility, for those polymers that are predominantly $\alpha$-(1\rightarrow3)-linked are insoluble whilst those that are predominantly $\alpha$-(1\rightarrow6)-linked are soluble (Walker, 1978; Rölla *et al.*, 1983; Walker & Jacques, 1987). The GTFs that produce these glucans are appropriately designated GTF-Is and GTF-Ss respectively.

To date, the sequences of seven streptococcal GTF-encoding genes have been published (Table 1). It is clear that all these genes are closely related and that their deduced amino acid sequences share a common domain structure in which the N-terminal two-thirds of the

Table 1. Oral streptococcal GTFs of known sequence

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Strain</th>
<th>Class</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GtfB</td>
<td><em>S. mutans</em> GS-5</td>
<td>GTF-I*</td>
<td>Shiroza <em>et al.</em> (1987)</td>
</tr>
<tr>
<td>GtfC</td>
<td><em>S. mutans</em> GS-5</td>
<td>GTF-SI*</td>
<td>Ueda <em>et al.</em> (1988)</td>
</tr>
<tr>
<td>GtfD</td>
<td><em>S. mutans</em> GS-5</td>
<td>GTF-S</td>
<td>Honda <em>et al.</em> (1990)</td>
</tr>
<tr>
<td>GtfI (dn)</td>
<td><em>S. downei</em> MFr28</td>
<td>GTF-I</td>
<td>Ferretti <em>et al.</em> (1987)</td>
</tr>
<tr>
<td></td>
<td>(formerly <em>S. sobrinus</em>)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GtfI (sb)</td>
<td><em>S. sobrinus</em> 6715</td>
<td>GTF-I</td>
<td>Abo <em>et al.</em> (1991)</td>
</tr>
<tr>
<td>GtfS</td>
<td><em>S. downei</em> MFr28</td>
<td>GTF-S</td>
<td>Gilmore <em>et al.</em> (1990)</td>
</tr>
<tr>
<td>GtfJ</td>
<td><em>S. salivarius</em> ATCC 25975</td>
<td>GTF-I</td>
<td>Giffard <em>et al.</em> (1991)</td>
</tr>
</tbody>
</table>

*These enzymes are reported to synthesize mixtures of soluble and insoluble polymers.

Abbreviation: GTF, glucosyltransferase.

The nucleotide sequence data reported in this paper have been submitted to GenBank and have been assigned the accession number Z11872 (gtfK and orfI).
enzyme, designated the 'catalytic domain', is responsible for the cleavage of sucrose (Ferretti et al., 1987). The C-terminal one-third, on the other hand, will bind to glucan and is known as the 'glucan-binding domain' (Ferretti et al., 1987; Mooser & Wong, 1988; Kobayashi et al., 1989). An active-site aspartic acid residue has been located in the catalytic domain (Mooser & Iwaoka, 1989; Mooser et al., 1991), while the glucan-binding domain is composed of a series of glycine-rich repeat sequences that exhibit similarity to the repeats in the glucan-binding protein from Streptococcus mutans (Banas et al., 1990) as well as the ligand-binding domains in Clostridium difficile toxin A and the lysins from Streptococcus pneumoniae (Wren, 1991).

In this communication, we report the sequencing of gfK, which encodes a primer-dependent GTF-S in S. salivarius ATCC 25975 (Pitty et al., 1989), and the comparison of the deduced amino acid sequence of GfK with seven other Gfts. A plausible multiple alignment of the central portion of the Gtf sequences was constructed and used to deduce a phylogenetic tree.

Methods

Strains and growth conditions. Plasmids were maintained in Escherichia coli strain NM522 (Gough & Murray, 1983) and derivatives of bacteriophage λ were grown on E. coli strain LE392 (Murray et al., 1977). All S. salivarius DNA was derived from J-A33, which has been described previously (Pitty et al., 1989; Giffard et al., 1991). E. coli cells were routinely grown on Luria Bertani (LB) medium (Miller, 1972) supplemented with ampicillin (100 μg ml⁻¹), isopropyl β-D-thiogalactoside (IPTG) (1 mM), or 5-bromo-4-chloro-3-indolyl β-D-galactoside (X-Gal) (100 μg ml⁻¹) as appropriate.

DNA sequence determination. All sequence determinations were carried out using the chromosome-termination method (Sanger et al., 1977). DNA that was to be sequenced was first subcloned into either or both of the pBR-containing plasmids pBl130 or pBlIII (IBI corporation), using routine procedures (Maniatis et al., 1982). The subcloned fragments are indicated in Fig. 1. In general, sequencing in the first direction was accomplished using a series of nested deletions constructed by the exonuclease III method of Henikoff (1984), and then confirmed in the other direction using custom-synthesized oligonucleotide primers. However, on occasion, sequencing was carried out in both directions using custom-synthesized primers. Templates were either single stranded or double stranded. Single-stranded templates were prepared using the helper phage M13K07 by the method recommended by Pharmacia for their pTZ series of plasmids, while double-stranded templates were prepared by the method of Birnboim & Doly (1979) and further purified by ultra-centrifugation through CsCl gradients (Maniatis et al., 1982). All sequencing reactions were carried out using the Pharmacia T7 sequencing kit according to the manufacturer's instructions. Complete sequences were obtained in both directions.

Sequences were assembled using the IBI Pustell sequence analysis software version 2.03. Database searches were carried out using the 'maf' program (Pearson & Lipman, 1988) in the Australian National Genome Information Service (ANGIS) based at the University of Sydney, Australia.

Multiple alignments and phylogenetic tree constructions. Both were carried out using the 'ClustalV' program (Higgins et al., 1992) in the ANGIS facility at the University of Sydney. The alignment parameters used are described in the appropriate sections. Tree construction was done using the neighbour-joining method (Saitou & Nei, 1987). In all cases, positions where there was a gap in any of the sequences were ignored, and branches were lengthened to account for multiple substitutions according to the algorithm of Kimura (1983). Confidence limits were calculated by means of the 'bootstrapping' method of Felsenstein (1985) using 1000 samples.

Divergence analyses. The divergences at silent sites (changes that do not alter the identity of the encoded amino acid) and replacement sites (changes that do alter the identity of the encoded amino acid) for pairs of coding sequences were calculated using the program 'Diverge' available within the Genetics Computer Group (GCG) collection of sequence analysis programs accessed through the ANGIS facility. This program makes use of the algorithm and the 'categories' of nucleotide changes described by Perler et al. (1980).

The input for the analyses were in-frame nucleic acid sequence alignments. The nucleic acid alignments were either derived from the amino acid sequence alignments, or were generated by aligning the nucleic acid sequences directly using the 'Best-Fit' program in the GCG package. The presence of gaps in the alignment, and also the necessity to discern local levels of divergence, required that the analyses be carried out using discrete blocks of alignment. The precise conditions used in each experiment are described in the relevant sections.

Results

The sequence of gfK and the downstream region

We have previously suggested that gfK, which encodes a primer-dependent GTF-S activity, lies 210 bp downstream from gfJ, which encodes a primer-dependent GTF-I activity, on the S. salivarius chromosome (Giffard et al., 1991; Fig. 1). As we had only reported the 5' end of the putative gfK gene, DNA sequence determination was carried out from the limit of this previously published sequence to the right-hand end of the cloned DNA (Fig. 1). It was found that the beginning of the putative gfK could be extended to give an open reading frame (ORF) of 4797 bp which encoded a polypeptide of M₁, 176466 which could be readily aligned with other GTF sequences (see below). The gfK was also preceded by putative ribosome-binding site and promoter elements (Fig. 2).

---

Fig. 1. Physical map of the gfJK locus, modified from Giffard et al. (1991). Filled areas represent the coding regions and the arrows the direction of transcription. The subclones used for sequencing gfK and orfI are also shown. H, HindIII; E, EcoR I; B, BglII; S, Sac I.
The C-terminal third of all sequenced GTFs is composed of a complex series of repeated sequences that have been shown to be associated with glucan-binding. These repeats have been divided into four classes (A–D) (Ferretti et al., 1987; Gilmore et al., 1990; Giffard et al., 1991) although the different classes do exhibit a certain level of similarity. Of these classes, the ‘A’ repeat has been found in multiple copies in all GTFs sequenced to date. The GtfK sequence was no exception as it contained seven ‘A’ repeats, of which five were immediately followed by sequences that showed similarity to the ‘D’ repeats of GtfJ and thus have been classified accordingly. The sequences of the repeats in GtfK and the location of the repeat classes in all known GTF sequences are shown in Fig. 3.

The GTFs are secreted enzymes and so would be expected to possess cleavable N-terminal signal sequences. Ferretti et al. (1987) have carried out N-terminal sequence analysis on the mature secreted form of Gtf-I from S. downei and have shown that a signal sequence cleavage site exists downstream of a hydrophobic region that follows von Heijne’s ‘−1, −3 rule’ (von Heijne, 1983). The GtfK sequence has an homologous hydrophobic stretch of amino acids followed by two alanine residues that occur at the ‘−1, −3’ sites in GtfI and so is also likely to be cleaved at this position (Fig. 4).

An ORF of unknown function follows gtfK

The sequenced DNA also defined another ORF 141 bp downstream from gtfK that we have designated orf1 (Fig. 1). This ORF is 345 bp in length and is preceded by putative ribosome-binding site and promoter elements (Fig. 2). It defines a highly hydrophobic polypeptide of M, 12,884. The deduced amino-acid sequence was used to search available sequence databases. While no examples of clear similarity were found, the sequence did show low levels of similarity with a number of membrane-associated proteins (data not shown), although by the criteria of von Heijne (1983) there was no clear evidence of any cleavable translocation signal sequence. The function of the product of orf1 remains unknown.

Multiple alignment and phylogenetic comparison of available Gtf sequences

In order to investigate the relationship between different Gtf sequences, it was first necessary to construct a plausible multiple alignment of a portion of the eight Gtf sequences. This was necessary as two regions within the eight sequences aligned extremely poorly (data not shown). Not unexpectedly, the first of these was the C-terminal glucan-binding domain with its apparent large number of duplications and rearrangements (Fig. 3). The second was a region of approximately 200 amino acids immediately following the signal sequence which was found to be highly variable and difficult to align reliably. Omission of both these domains from the ClustalV analysis yielded a multiple alignment of the eight Gtf sequences that contained a relatively small number of gaps. In order to obtain the ‘best’ alignment possible, the gap penalties were progressively modified. Significant adjustment of the pairwise gap penalty had no effect on the results, so attention was given to adjusting the multiple alignment fixed gap penalty and floating gap penalty, which control the opening of a gap and lengthening of a gap respectively. Progressively altering these gap penalties had only a slight effect on the multiple alignment until certain values were reached, at which point essential gaps were not opened and the multiple alignment became implausible.

Different alignments were tested by building phylogenetic trees using the neighbour-joining method (Saitou & Nei, 1987) and comparing the bootstrapping confidence limits, topologies and branch lengths. In all plausible alignments, the confidence limits were 100% for all branch orders except for the shortest, which varied between 75% and 90%, the topologies were all identical, and branch lengths the same within 5%. By subjective judgment, the alignment obtained by using a fixed gap penalty of 10 and a floating gap penalty of 30 was chosen for further analysis (Fig. 5).

A phylogenetic tree was constructed from this alignment excluding all positions where gaps occurred (Fig. 6). The topology was tested by bootstrapping (Felsenstein, 1985) using 1000 samples. All the branching orders gave a confidence limit of 100% except for the separation of GtfJ, GtfK and GtfD from the other sequences which gave a confidence limit of 79.3%.

Two significant observations could be drawn from the phylogenetic analysis shown in Fig. 6. First, if one excluded the groupings of GtfB/GtfC, and the two GtfIs (both of which are identical over much of their lengths – see below), there was no clustering of enzymes that make similar glucan products. Neither GtfI and the two GtfIs, which make insoluble products, nor GtfS, GtfD and GtfK, which make soluble products, form natural groups. Since insoluble glucans are primarily α-(1→3)-linked while soluble glucans are α-(1→6)-linked (Walker & Jacques, 1987; Pitty et al., 1989; Gilmore et al., 1990), it can be inferred that different protein linkage specificities may have evolved more than once. Secondly, there was little difference between the maximum divergence between sequences from within the mutans group and the divergence between the sequences from S. salivarius and the mutans group. This indicated that in the absence of inter-species gene transfer, the existence of multiple
Fig. 2. Nucleotide sequences and deduced amino-acid sequences of gtfA and orfJ. Potential ribosome binding sites (RBS) and −10 and −35 promoter elements are indicated.

and the mutants group. This indicated that in the absence of inter-species gene transfer, the existence of multiple GTFs might pre-date the divergence of the mutants group from S. salivarius.

Evidence for horizontal gene transfer
Partial recombination events will result in different parts...
defined classes of repeats in the C-terminal glucan-binding regions are shown in (b). GBP, glucan-binding protein (S. mutans).

Fig. 3. Repeated sequences from the C-terminal domain of GtfK (a) compared with consensus sequences (Con.) of the ‘A’ repeat (Gilmore et al., 1990) and ‘D’ repeat (Giffard et al., 1991). The locations of defined classes of repeats in the C-terminal glucan-binding regions are shown in (b). GBP, glucan-binding protein (S. mutans).

Fig. 4. Comparison of the N-terminal cleavable signal sequence of the GtfI from S. downei (Ferretti et al., 1987) with the N-terminus of GtfK. The GtfI has alanine residues at –1 and –3 with respect to the cleavage site while GtfK possesses an homologous site.

of the same gene having different evolutionary histories. In order to test for evidence of such past partial recombination events the region used to constitute the multiple alignment in Fig. 5 was divided into seven blocks of 100 positions and one block of 129 positions (positions where there was a gap in any of the sequences were not counted), and these blocks (Fig. 5) used to construct phylogenetic ‘sub-trees’ as outlined above. These sub-trees were then examined for localized changes in topology such as the existence of significant groupings in the sub-trees that did not exist in the overall tree (Fig. 6), or proportionally large changes in the distance between any two Gtfs.

From the eight sub-trees obtained (Fig. 7) several observations were made. First, the distance between GtfB and GtfC decreased dramatically over the length of the sequence analysed. In particular, there was a relatively high divergence between the two sequences in the first two sub-trees, but in the last six sub-trees the sequences were almost identical. This suggested that there had been a recent partial recombination between gtfB and gtfC that involved sequences distal to block 1. This was plausible since gtfB and gtfC lie next to each other on the S. mutans GS-5 chromosome (Hanada & Kuramitsu, 1988). Secondly, in the phylogenetic tree (Fig. 6), GtfJ and GtfK form a significant group. However, this was only the case in sub-trees 1, 3, and 5 (Fig. 7), although in no sub-trees did the sequences come particularly close. This suggested that there may have been a recombination between gtfJ and gtfK, but not as recently as between gtfB and gtfC. Again this was plausible because gtfJ and gtfK, like gtfB and gtfC, lie next to each other on the chromosome (Giffard et al., 1991). Thirdly, although the GtfI sequences from S. downei and S. sobrinus showed a high level of similarity throughout the analysed region, there did appear to be proportionally large local changes in the divergence between these two sequences. In particular, the divergence between these two sequences was much greater in sub-trees 1 and 3 than in the other sub-trees (Fig. 7). This finding was somewhat unexpected because S. downei (formerly S. mutans serotype h, then S. sobrinus serotype h) is a monkey isolate (Beighton et al., 1981) while S. sobrinus 6715 (formerly S. mutans 6715) is a human isolate (Fitzgerald et al., 1968), making recombination
Fig. 5. Multiple alignment of the central portion of the deduced amino acid sequences of the catalytic domains of the eight available Gtfs. The constraints used in constructing the alignment are described in the text.

The main blocks (BLOCK 1–BLOCK 8) as well as the three sub-blocks (BLOCK A–BLOCK C) used in subsequent phylogenetic analyses are shown. The residues aligned are: GtfJ, 180–1108; GtfK, 176–1078; GtfB, 166–1037; GtfC, 191–1075; GtfI (dn), 168–1046; GtfI (sb), 162–1035; GtfS, 156–1026; GtfD, 166–1069.

somewhat unlikely. Fourthly, in the overall phylogenetic tree (Fig. 6) GtfB and GtfC formed a significant group with the two GtfIs. However, in sub-tree 6, both GtfB and GtfC no longer formed a significant group with the two GtfI sequences but instead formed a grouping with GtfD with borderline significance. This suggested that both GtfB and GtfC contain sequences that originated from GtfD, implying a small partial recombination between gtfD and the precursor of gtfB/gtfC. Again this was plausible because gtfD is found in the same species as gtfB and gtfC. Lastly, in no case did GtfS, the GtfI's or GtfJ and GtfK show any sign of forming significant
groupings in the sub-trees (Fig. 7) that did not exist in the overall phylogenetic tree (Fig. 6).

**Silent-site divergence analysis**

The possibility of partial recombination events between \textit{gtfB} and \textit{gtfC} and between \textit{gtfJ} and \textit{gtfK} as investigated further using the replacement-site and silent-site divergence algorithms of Perler et al. (1980). This strategy depends on the principle that nucleotide identities at codon ‘wobble’ positions (or silent sites) are not constrained by the functional requirements of the gene product, so if two related but un-recombined genes are compared, they will exhibit a relatively constant level of silent-site divergence over their entire lengths. On the other hand, a recent partial recombination event will result in localized changes in the level of silent-site divergence. These analyses were carried out by determining the nucleotide sequences for blocks of pairwise aligned amino acid sequences within the entire catalytic domain and measuring the divergence of these nucleotide sequences at silent sites and replacement sites. Where possible, the blocks were arbitrarily made 99 nucleotides in length. However, the presence of gaps in the alignments necessitated some flexibility in block size, giving rise to blocks of between 60 and 150 nucleotides in length. In the case of \textit{gtfB} and \textit{gtfC}, it was clear that the divergence at silent sites was very high at the beginning of the genes, but then decreased significantly around nucleotide 1800 (Fig. 8a), strongly supporting the contention that these two genes had undergone a partial recombination. In the case of \textit{gtfJ} and \textit{gtfK} (Fig. 8b), the situation was different. There was no clear evidence of conservation at silent sites, although there may have been slight conservation in the region between residues 500 and 1000. This region coincides with the first of the sub-trees subjected to phylogenetic analysis, in which \textit{GtfJ} and \textit{GtfK} formed a highly significant group. Thus while there was some internally consistent evidence for partial recombination between \textit{gtfJ} and \textit{gtfK} it was far from clear-cut, indicating that such an event, if it had taken place at all, had not happened recently.

A similar approach was applied to the comparison of the two \textit{gtfI} sequences, although in this case the very close relationship between the \textit{GtfIs} meant that the nucleotide sequences could be readily aligned and much larger blocks of sequence used directly as input into the ‘Diverge’ program. It is clear from this analysis that there were dramatic differences in the silent-site divergence along the length of the two genes (Table 2). This observation supported the notion that there had been a partial recombination between these two sequences, despite their isolation from different bacterial species with different hosts. Also of interest was the large difference between the values of the silent-site and replacement-site divergences despite the high degree of sequence similarity. This would seem to indicate that the majority of the residues within these two enzymes can only undergo a small number of possible changes in order for these changes to remain selectively neutral.

Lastly, this approach was applied to a comparison of \textit{GtfD} with \textit{GtfB} and \textit{GtfC}. In practice, \textit{GtfB} was excluded from this analysis because of the almost perfect identity of \textit{GtfB} and \textit{GtfC} in the analysed region, and also because of a localized reading frame discontinuity in the alignment of \textit{GtfB} with the other sequences observed.

**Table 2. Silent-site and replacement-site divergence within the catalytic domains of the GtJls of \textit{S. downei} and \textit{S. sobrinus}**

<table>
<thead>
<tr>
<th>Nucleotides from start of coding region</th>
<th>Corrected percentage divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent sites</td>
<td>Replacement sites</td>
</tr>
<tr>
<td>1–248</td>
<td>69.4</td>
</tr>
<tr>
<td>267–1835</td>
<td>64.0</td>
</tr>
<tr>
<td>1836–2987</td>
<td>3.2</td>
</tr>
</tbody>
</table>

**Table 3. Silent-site and replacement-site divergence within blocks A, B and C of the GtfC and GtfD of \textit{S. mutans}**

<table>
<thead>
<tr>
<th>Blocks of alignment (see Fig. 5)</th>
<th>Corrected percentage divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silent sites</td>
<td>Replacement sites</td>
</tr>
<tr>
<td>Block A</td>
<td>231.0</td>
</tr>
<tr>
<td>Block B</td>
<td>194.1</td>
</tr>
<tr>
<td>Block C</td>
<td>170.2</td>
</tr>
</tbody>
</table>
Evolution of gtf genes

Fig. 7. Phylogenetic 'sub-trees' 1–8 (a–h) derived from blocks 1–8, respectively, of the alignment in Fig. 5. Bootstrapping confidence limits over 90% are shown.

during the course of this analysis. [This discontinuity encompasses GtfB nucleotides 2397–2450 (numbered from the start codon) and may be explained by either a very recent single nucleotide deletion/insertion event in gtfB, or an error in the reported sequence of gtfB.] Furthermore, because of the overall relatively high level of divergence of GtfD from GtfB/GtfC, the analysis was confined to the region of interest highlighted by sub-tree analysis of block 6 (Fig. 7f). Three blocks of 150 nucleotides were considered. The central block (block B, Fig. 5) was judged to encompass the amino acid sequences that resulted in the grouping of GtfD with GtfB and GtfC (Fig. 7f), while the two other blocks (block A and block C, Fig. 5) occurred immediately before and after block B and were included for comparison. It was evident from the levels of silent-site divergence in all three blocks that there was no sign of any conservation at silent sites; in fact the analyses indicated a saturating level of substitutions in all three blocks (Table 3). Consequently, if recombination between GtfD and the precursor of GtfB and GtfC had occurred in this region, it was too long ago for conservation at silent sites to be preserved.

Lack of evidence for inter-species transfer of intact gtf genes

The most common evidence presented for interspecies gene transfer is the presence of closely related genes in unrelated backgrounds, sometimes supported by base composition studies. However, in the case of the gtf genes, all the species are fairly closely related and have
similar base compositions. Therefore, the only approach open to us was to compare the tree in Fig. 6 with the phylogenies derived by DNA hybridization and rRNA sequencing (Schleifer et al., 1984; Ludwig et al., 1985; Schleifer & Kilpper-Balz, 1987; Bentley, 1991). These latter analyses have consistently found that the mutans group streptococci form a phylogenetically valid (albeit somewhat loose) grouping that does not include $S$. salivarius, that the extent of divergence within the mutans group is not greatly different from the divergence between the mutans group and $S$. salivarius, and that there is a specific close relationship between $S$. downei and $S$. sobrinus. If it were assumed that GtfB, GtfC, and the two Gtfs are all homologues in the strictest sense (i.e. are the products of the same gene duplication/species divergence events), and one does not consider GtfS and GtfD, then the phylogenies derived from DNA hybridization and rRNA sequences are fully consistent with the Gtf phylogenetic tree in Fig. 6. Specifically, the GTF-1 enzymes from $S$. mutans and $S$. downei/$S$. sobrinus form a fairly loose but valid group that does not include sequences of $S$. salivarius origin, and the Gtfs from $S$. downei and $S$. sobrinus clearly cluster together. Since GtfS and GtfD do not form a group they remain something of a puzzle. It may be that they are not strict homologues and therefore may be the products of different duplication events, having diverged prior to $S$. mutans and $S$. downei becoming separate species. Alternatively, there may be somewhat different selection pressures on GTF-S enzymes, resulting in faster divergence. Whatever the interpretation, there are no groupings in the Gtf phylogenetic tree that do not exist in the rRNA tree, implying that there is no evidence for the transfer of intact gtf genes between the species so far studied.

**Discussion**

The sequencing of gtfK completes the sequence analysis of the gtfJK locus as reported by Giffard et al. (1991). The arrangement and spacing of the genes is reminiscent of the gtfB and gtfC of $S$. mutans GS-5, although the gene products have somewhat different functions. Whereas both GtfB and GtfC have been reported to be primer-independent enzymes that synthesize a mixture of soluble and insoluble glucans (Aoki et al., 1986; Hanada & Kuramitsu, 1988), GtfJ and GtfK are both primer-dependent enzymes that synthesize insoluble and soluble glucans respectively (Pitty et al., 1989). Thus, the activity of GtfJ most resembles that of the Gtfs of $S$. downei and $S$. sobrinus while that of GtfK is similar to the GtfD of $S$. mutans.

GtfK can be clearly aligned with all other Gtf sequences, and contains the putative signal sequence and C-terminal repeats that are typical of this family of enzymes. As with the other Gtfs, the most abundant repeated sequence is the ‘A’ repeat while five of these ‘A’ repeats are followed by another class of repeat that shows similarity to the ‘D’ repeats of GtfJ and therefore has been classified as such. However, the classification of these repeats represents something of a problem in our view, as all of them show a certain level of similarity with one another as well as to the repeats in the ligand-binding domains of the toxins from Clostridium difficile and in the lysins from Streptococcus pneumoniae (Wren, 1991), which are classified according to a completely different scheme. The possibility of a common underlying structure to all these repeat classes is currently under investigation.

Our evolutionary analysis of the eight gtf genes so far sequenced was based on constructing a phylogenetic tree from a highly homologous region in the catalytic domain of these enzymes and then attempting to falsify it by searching for evidence of recombination and/or horizontal gene transfer. The lengths of the sequences aligned meant that a phylogenetic tree with extremely good confidence limits could be constructed. This analysis showed that enzymes with similar activities did not group together and that the existence of multiple Gtfs
may have pre-dated the divergence of *S. salivarius* from the mutants group.

Partial recombination events were tested for by seeking out localized deviations from the overall relationships between the sequences by constructing a series of subtrees from portions of the alignment and also by carrying out a series of silent-site divergence measurements. From these analyses, clear corroborative evidence was obtained for the partial recombination between *gtfB* and *gtfC*, and also between the *gtfI* genes of *S. downei* and *S. sobrinus*. Recombination between *gtfB* and *gtfC* was not surprising as these two genes lie next to each other on the *S. mutans* GS-5 chromosome. Furthermore, polymorphisms at this locus have been documented, indicating the potential for rearrangements between these two genes (Chia et al., 1991; Yamashita et al., 1992). In contrast, however, the apparent partial recombination between the two *gtfI* genes was unexpected. These genes are not only found in different species of streptococci, but the streptococci themselves have different hosts. Gene transfer between the two species would therefore appear at first sight to be unlikely. It may simply be that the two species live in both monkeys and man. Whatever the explanation, these recombination events imply that the distances between GtfB and GtfC and between the two GtfIs in Fig. 6 are an underestimation of the true phylogenetic distances.

The evidence for recombination between the precursor of *gtfB* and *gtfC* with *gtfD* and also for recombination between *gtfJ* and *gtfK* was much less convincing because it could not be corroborated by silent-site divergence analysis. However, because each group of genes is resident on the chromosome of a specific species, recombination events would not be unexpected. Thus while recombination is probable it remains unconfirmed, and so the divergence between the sequences can be thought of as a minimum not a maximum relative to the time since the sequences diverged.

Apart from the situation with the two *gtfI* genes, it was difficult to discern any evidence for inter-species gene transfer. However, such events are very difficult to rule out. GTF enzymes are by no means ubiquitous in the genus *Streptococcus*, yet rRNA sequence studies (Bentley et al., 1991) show no evidence that the mutants group and *S. salivarius* form any sort of natural group. As both species contain *gtf* genes it is plausible that an ancestral *gtf* gene(s) was transferred from the mutants group common ancestor to the *S. salivarius* ancestor or vice versa. However, it is not necessary to invoke such an event to rationalize the rRNA phylogenetic tree with the Gtf phylogenetic tree since the Gtf divergence evident both within the mutants group and within *S. salivarius* suggests that such an event must be very ancient if it happened at all. The unravelling of these early events must await a more sophisticated analysis.

This work was supported by a Program Grant awarded by the Australian National Health and Medical Research Council and in part by the Australian Dental Research Fund Inc.

**References**


