Genomic Organization and Molecular Analysis of Virulent Bacteriophage 2972 Infected an Exopolysaccharide-Producing Streptococcus thermophilus Strain

Céline Lévesque,1† Martin Duplessis,1,2 Jessica Labonté,1,2 Steve Labrie,1,2 Christophe Fremaux,4 Denise Tremblay,1,3 and Sylvain Moineau1,2,3,*

Groupe de Recherche en Écologie Buccale (GREB), Faculté de Médecine Dentaire,1 Département de Biochimie et de Microbiologie, Faculté des Sciences et de Génie,2 and Centre de Référence pour Virus Bactériens Félix d’Hérelle,3 Université Laval, Québec, Canada G1K 7P4, and Danisco, BP10, 86220 Dangé Saint Romain, France*

Received 4 November 2004/Accepted 1 February 2005

The Streptococcus thermophilus virulent pac-type phage 2972 was isolated from a yogurt made in France in 1999. It is a representative of several phages that have emerged with the industrial use of the exopolysaccharide-producing S. thermophilus strain RD534. The genome of phage 2972 has 34,704 bp with an overall G+C content of 40.15%, making it the shortest S. thermophilus phage genome analyzed so far. Forty-four open reading frames (ORFs) encoding putative proteins of 40 or more amino acids were identified, and bioinformatic analyses led to the assignment of putative functions to 23 ORFs. Comparative genomic analysis of phage 2972 with the six other sequenced S. thermophilus phage genomes confirmed that the replication module is conserved and that cos- and pac-type phages have distinct structural and packaging genes. Two group I introns were identified in the genome of 2972. They interrupted the genes coding for the putative endolysin and the terminase large subunit. Phage mRNA splicing was demonstrated for both introns, and the secondary structures were predicted. Eight structural proteins were also identified by N-terminal sequencing and/or matrix-assisted laser desorption ionization—time-of-flight mass spectrometry. Detailed analysis of the putative minor tail proteins ORF19 and ORF21 as well as the putative receptor-binding protein ORF20 showed the following interesting features: (i) ORF19 is a hybrid protein, because it displays significant identity with both pac- and cos-type phages; (ii) ORF20 is unique; and (iii) a protein similar to ORF21 of 2972 was also found in the structure of the cos-type phage DT1, indicating that this structural protein is present in both S. thermophilus phage groups. The implications of these findings for phage classification are discussed.

Streptococcus thermophilus is one of the most economically important lactic acid bacteria (LAB) used for the manufacture of yogurt and Swiss- or Italian-type hard cooked cheeses (19). This bacterium may also play a role as a probiotic, alleviating symptoms of lactose intolerance and other gastrointestinal disorders (28). Research on the physiology of S. thermophilus has generated significant insights into some of its properties, including sugar metabolism, protein utilization, and exopolysaccharide (EPS) production (7, 19). S. thermophilus belongs to the group of bacteria that are generally recognized as safe, which is an exception in the genus Streptococcus.

S. thermophilus bacteriophages have been a subject of ongoing interest, because they are ubiquitous in dairy environments and because their rapid lytic cycle can lead to significant bacterial lysis that results in milk fermentation delays (52). Many strategies have been employed by dairy factories to curtail phage infections. One extensively used tactic is the rotation of several LAB strains to prevent the proliferation of specific phage populations. Additionally, carefully selected so-called phage-insensitive S. thermophilus strains are introduced into dairy processes with the hope of limiting phage infections (49). However, despite these efforts, new S. thermophilus phages are still emerging. It is expected that the characterization of an increasing number of streptococcal phage genomes should lead to a better understanding of phage evolution, which is required for the development of long-term phage-resistant LAB strains.

S. thermophilus phages are a relatively homogeneous group with the same morphology (B1 morphotype, Siphoviridae family) (1). They have an isometric capsid (diameter, 45 to 60 nm) and a long, noncontractile tail of various lengths (240 to 270 nm) and thicknesses (9 to 13 nm) (8). They are divided into two groups based on the packaging mechanism of their double-stranded DNA (cos and pac types) and the number of major structural proteins (37). Six complete genome sequences of S. thermophilus phages are currently available. They include the cos-type phages DT1 (66), Sfi19 (42), Sfi21 (12), and 7201 (62), as well as the pac-type phages O1205 (61) and Sfi11 (39, 40). DT1, Sfi19, and Sfi11 are virulent phages, while the others are temperate.

Comparative genomic analyses of these six genomes demonstrated that S. thermophilus phages share extensive DNA sequence similarity in the replication module and lysis cassette. Significant differences have been reported in the genes coding for structural proteins, which is in agreement with the classification scheme (20, 37). An interesting feature is the close genetic relationship between virulent and temperate S. thermophilus phages. It has even been proposed that virulent S.
therophilus phages arose from temperate phages through a combination of rearrangement and deletion events within the lysogenic module (11, 41).

One of the most significant contributions of the streptococcal phage genomic analyses has been in the field of phage taxonomy. These comparative analyses revealed the presence of related phages in other species and genera of low-G+C-content gram-positive bacteria (9). Another benefit of these genomic studies has been the use of some phage genetic elements to construct antiphage systems. These elements include the phage origin of replication (26, 62, 63), the CI-like repressor (14), the immunity gene (13), and the antisense RNA technology targeting the putative helicase and primase genes of S. thermophilus phages (63, 64).

In the present work, we report the complete nucleotide sequence and molecular characterization of 2972, a virulent pac-type phage that infects the exopolysaccharide-producing strain S. thermophilus RD534, which is used for the production of yogurt worldwide.

**MATERIALS AND METHODS**

**Phage preparation and purification.** The virulent S. thermophilus phages infecting strain S. thermophilus RD534 were provided by Danisco (France). For phage propagation, S. thermophilus RD534 was grown at 42°C without agitation in M17 broth (Quélab, Québec, Canada) supplemented with 0.5% (wt/vol) lactose and 10 mM CaCl₂. When the optical density at 600 nm reached 0.2, approximately 10⁷ PFU/ml of phage was added and the culture was incubated overnight at 42°C. The lysate was clarified by centrifugation and passed through a 0.45-µm pore-size filter. Phages were purified by ultracentrifugation using a discontinuous CsCl gradient (56). Phage morphology was observed as described previously (50) with a Philips 420 transmission electron microscope operating at 80 kV.

**Purification of phage DNA and DNA sequencing.** Phage DNA was isolated using the QIAGEN lambda Maxi kit as described previously (31). DNA restriction profiles were analyzed using Molecular Analyst Fingerprinting Plus software (Bio-Rad Laboratories) and compared using the UPGMA (unweighted-pair group method using average linkages) clustering method. Phage 2972 DNA was sequenced from shotgun subclone libraries of the genome (Integrated Genomics, Inc., Chicago, IL). Then the gap between contigs was closed by sequencing gap-specific PCR products generated by using phage 2972 genomic DNA as a template; this procedure was performed by the DNA sequencing service of Université Laval. Computer-assisted DNA and protein analyses were performed using the Genetics Computer Group Sequence Analysis software package, version 10.3 (22). The genome sequence was analyzed using the open reading frame (ORF) finder, graphical analysis tool (http://www.ncbi.nlm.nih.gov/000101.html) to define potential coding regions. The PROSITE and Pfam databases were employed (http://hits.isb-sib.ch/cgi-bin/PFSCAN) to locate putative functional motifs. PSI-BLAST and Advanced BLAST Search 2.1 (http://www.ncbi.nlm.nih.gov/BLAST) were also used for sequence comparisons with databases (2).

**RNA methods.** Total RNA was extracted from phage 2972-infected S. thermolus cells (17 min after infection) to study mRNA splicing. Transcription was stopped by adding rifampin at 150 µg per ml. Cells were collected by centrifugation and frozen in dry-ice-ethanol. The frozen cell pellets were resuspended in 1 ml of TRIZOL reagent (Invitrogen) and transferred to a 2-ml tube containing 0.7 g of glass beads (106 µm; Sigma-Aldrich). The mixture was vortexed with a Mini-Beadbeater-8 cell (BioSpec Products) three times, for 1 min each time (67). Between treatments, the cell suspensions were chilled on ice for 1 min. The supernatant was extracted twice with TRIZOL-chloroform. Nucleic acids were precipitated with isopropanol and resuspended in diethyl pyrocarbodine (DEPC)-treated water. Samples were treated with DNase I (10 U) for 30 min at 37°C with 80 U of RNaseOUT recombinant RNase inhibitor (Invitrogen).

DNA-free RNA samples were subjected to reverse transcription (RT) as follows. Ten micrograms of purified RNA and 6 µg of oligonucleotides (random hexamers; Invitrogen) were added to DEPC-treated water to obtain a final volume of 18.5 µl. The mixture was heated at 70°C for 10 min and snap-frozen in dry-ice-ethanol for 30 s. Then 400 U of SuperScript II RNase H⁻ reverse transcriptase (Invitrogen), 6 µl of 5× First Strand buffer (Invitrogen), 3 µl of 0.1 M dithiothreitol (DTT) (Invitrogen), and 0.75 mM dATP, dCTP, dGTP, and dTTP were added to the mixture, and the RT reaction was performed at 42°C for 16 h. The reaction was terminated by heating at 75°C for 15 min. The cDNA was amplified by PCR as described previously (23). The cDNA was heated at 94°C for 4 min, followed by 35 cycles of the following temperature-time profile: 94°C for 45 s, 57°C for 45 s, and 73°C for 1 min. After the final cycle, the mixtures were kept at 73°C for an extra 10 min. The primers used for amplification of the gene encoding the large subunit of the terminase were 5’-CTATCAAAGCAAGCTAC GCCC-3’ (forward) and 5’-CTCTACCCAGACTACCAAGATA3’ (reverse), and the primers used for the endolysin-encoding gene were 5’-GAATTCCAAATAC GTAAACGG-3’ (forward) and 5’-CTTCAAGTTGCCATCTGGAGA3’ (reverse). PCR products were separated by electrophoresis on agarose gels (2%), stained with ethidium bromide, and visualized by UV. PCR products were also purified using QIAquick PCR purification columns (QIAGEN) and sequenced with the same primers used for the PCR amplification.

**Phage structural protein analysis.** Phage 2972 structural proteins were separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) with a 15% polyacrylamide separating gel and a 4.5% polyacrylamide stacking gel (56). The proteins were then transferred to a polyvinylidene difluoride Immobilon-P PVDF membrane (Millipore). After staining with 0.1% (wt/vol) Coomassie blue in 40% (vol/vol) methanol and 1% (vol/vol) acetic acid, the protein bands of interest were excised. N-terminal sequencing was performed by Edman degradation using an Applied Biosystems model 473A pulsed liquid protein sequencer. For matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) mass spectrometry, Coomassie-stained proteins were extracted from the polyacrylamide gels and digested with trypsin using a MassPrep liquid handling station (Micromass Ltd.) according to the manufacturer’s specifications. The resulting peptides were lyophilized and resuspended in 0.1% (vol/vol) formic acid (TFA) and subjected to 1:100 acetic acid (TFA)-acetonitrile (0.1% TFA) was used as a matrix for the MALDI analysis. Equal volumes of peptide and matrix solution were mixed and spotted onto a stainless-steel MALDI sample plate. The sample-matrix solution was allowed to air dry at room temperature and was then washed with 0.1% TFA. MALDI-TOF spectra were acquired on a Voyager-DE PRO Biospectrometry Workstation (Applied Biosystems) and analyzed using DataExplorer software, version 4.0 (Applied Biosystems). The instrument was operated in the positive-ion reflector delayed-extraction mode. The PeptIdent tool (http://ca.expasy.org/tools/peptident.html) was used to search nonredundant Swiss-Prot/TrEMBL protein databases for matching peptide mass fingerprints and to identify proteins. Search criteria allowed a maximum of one missed cleavage by trypsin, complete carboxy-amidomethylation of cysteine, partial methionine oxidation, and mass deviations under 60 ppm. The N-terminal sequencing and mass spectrometry analyses were both performed at the Eastern Québec Proteomic Center (Québec, Canada).

**Nucleotide sequence accession number.** The nucleotide sequence of the phage 2972 genome has been deposited in GenBank under accession no. AY699705.

**RESULTS**

**Phages infecting S. thermophilus RD534.** S. thermophilus RD534 is widely used for the commercial manufacture of yogurt because it produces an EPS that gives a viscous texture to the fermented dairy product. Compositional and structural analyses have revealed that the EPS produced by RD534 is composed primarily of β-glucose and β-galactose in a molar ratio of 1:1 (data not shown). The EPS of RD534 is thus similar to those produced by other S. thermophilus strains (38, 44). S. thermophilus RD534 is sensitive to a group of closely related virulent S. thermophilus phages (Fig. 1), all of which belong to the S. thermophilus pac-type group, as submolar fragments were still observed following the heating of restriction endonuclease digests of the phage DNA (data not shown). Phage 2972 was randomly selected as a representative of this group for further analysis. Electron microscopic analysis of the purified preparation of phage 2972 showed that it possesses a 55-nm-diameter isometric capsid and a 260-nm-long noncontractile tail (Fig. 2).

**Determination of the complete nucleotide sequence.** The genome of phage 2972 has 34,704 bp. It is the shortest S. thermophilus phage genome for which the complete nucleotide sequence is available. The genome of phage 2972 is 5.1 kb shorter than the genome of phage Sfi11 (39,807 bp), the other
virulent pac-type *S. thermophilus* phage for which the entire nucleotide sequence is available (39, 40). Phage 2972 DNA has an average G+C content of 40.15%, which is similar to the G+C content reported for the host genome (37.8 to 40%) (43). Forty-four ORFs of 40 codons or more were identified (Table 1). They were all located on the same strand and all started with either an ATG or a TTG initiation codon. Every ORF, except orf4 and orf37, was preceded by a region that shares variable homologies with the Shine-Dalgarno sequence complementary to the 3’ end of the 16S rRNA of *S. thermophilus* (AAAGGAGGTGA). Of the 44 ORFs, 23 could be assigned a putative function based on their similarity to proteins with known functions or on conserved motifs (Table 1). Three putative promoters (P1, P2, and P3) were identified by their similarity to the consensus −35 and −10 sequences. The P1 promoter (gTGATA–N16–TATAAT; lowercase letters indicate a difference from the consensus sequence) was located 358 bp upstream of the ATG start codon, and the P2 promoter (TTGACA–N20–TAc) was located 30 bp upstream of the orf31 start codon, and the P3 promoter (TTGACA–N20–TAcTt) was located 166 bp upstream of the orf39 start codon. Three putative terminator-like structures (factor-independent terminators) were also identified. Terminator T1 was located 18 bp downstream of the stop codon of orf1, T2 was located 64 bp upstream of the putative P2 promoter, and T3 was located 165 bp downstream of the orf38 stop codon. A 250-bp noncoding region was identified between orf38 and orf39. It contains four inverted repeats (75% A+T content within 140 bp) that may correspond to the origin of replication (ori) of the phage genome (6, 26). In all the *S. thermophilus* phage genomes analyzed to date, the phage ori is located just upstream of the genes coding for proteins involved in DNA replication.

### Analysis and organization of the genome

The genome of phage 2972 is organized into distinct modular regions commonly found in other phages of the *Siphoviridae* family (9, 20, 32). As can be seen in Fig. 3, two regions are highly conserved in the seven *S. thermophilus* phage genomes, including the segments containing the genes necessary for DNA replication and host cell lysis. One notable exception is the putative holin-encoding gene (orf25) in the lysis cassette of phage 2972. orf25 codes for a 108-amino-acid (108-aa) protein with significant similarities (>50% identity) with the holins of a *Streptococcus pyogenes* prophage 315.5 prophage (5), the *Streptococcus mitis* temperate phage SM1 (59), and the *Streptococcus pneumoniae* temperate phage MM1 (53) (Table 1). Topology prediction analyses identified three transmembrane domains in ORF25, a characteristic of type I holins (68). In most phages of the *Siphoviridae* family, the gene immediately downstream from the holin-encoding gene codes for the endolysin. In phage 2972, the products of orf26 and orf29 are both homologous (>76% identity) to the bacteriophage peptidoglycan hydrolases (amidases) of *S. thermophilus* phages DT1, 7201, Sfi19, Sfi21, Sfi11, and O1205 (Table 1). These two ORFs are separated by an intron. Another intron has been found in the phage 2972 genome. It is located between orf3 and orf4. The orf3 and orf4 genes, encoding proteins of 223 aa and 195 aa, respectively, exhibit high identity with the large subunit of the terminase of *S. thermophilus* phage Sfi11 (Table 1). The introns of phage 2972 are described below.

### Lysogeny module

In the genome of *S. thermophilus* temperate phages, the conserved region containing the DNA replication module is separated from the lysis cassette by the lysogeny module (Fig. 3). As in the virulent *S. thermophilus* phages DT1, Sfi11, and Sfi19, remnants of a lysogeny module are found in the genome of the virulent phage 2972. For example, orf31 likely codes for a *cro*-like repressor. It is possible that phage 2972 picked up DNA by homologous recombination with a...
<table>
<thead>
<tr>
<th>ORF</th>
<th>Start</th>
<th>Stop</th>
<th>Size (aa)</th>
<th>Mol mass (kDa)</th>
<th>pI</th>
<th>Putative function and motif&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Best match(es) (% amino acid identity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>414</td>
<td>827</td>
<td>137</td>
<td>16.1</td>
<td>8.7</td>
<td>—</td>
<td>ORF137 S. thermophilus phage Shi11 (135/137; 98%)</td>
</tr>
<tr>
<td>2</td>
<td>1009</td>
<td>1461</td>
<td>150</td>
<td>16.7</td>
<td>8.0</td>
<td>Small terminase</td>
<td>ORF25 S. thermophilus phage O1205 (74/147; 50%)</td>
</tr>
<tr>
<td>3</td>
<td>1448</td>
<td>2119</td>
<td>225</td>
<td>25.3</td>
<td>9.0</td>
<td>Large terminase</td>
<td>ORF411 phage Shi11 (211/217; 97%)</td>
</tr>
<tr>
<td>4</td>
<td>2404</td>
<td>2991</td>
<td>261</td>
<td>22.6</td>
<td>9.2</td>
<td>Large terminase</td>
<td>ORF411 phage Shi11 (183/192; 95%)</td>
</tr>
<tr>
<td>5</td>
<td>3000</td>
<td>4505</td>
<td>150</td>
<td>17.4</td>
<td>10.3</td>
<td>Portal protein</td>
<td>ORF27 phage O1205 (472/501; 94%)</td>
</tr>
<tr>
<td>6</td>
<td>4502</td>
<td>5395</td>
<td>90</td>
<td>13.8</td>
<td>10.3</td>
<td>Capsid protein</td>
<td>ORF28 phage O1205 (286/297; 96%)</td>
</tr>
<tr>
<td>7</td>
<td>5583</td>
<td>6164</td>
<td>70</td>
<td>9.2</td>
<td>10.3</td>
<td>Capsid protein</td>
<td>ORF29 phage O1205 (184/193; 95%)</td>
</tr>
<tr>
<td>8</td>
<td>6184</td>
<td>6543</td>
<td>119</td>
<td>12.7</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF319 phage Shi11 (110/119; 92%)</td>
</tr>
<tr>
<td>9</td>
<td>6562</td>
<td>7608</td>
<td>105</td>
<td>13.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF348 phage Shi11 (326/348; 93%)</td>
</tr>
<tr>
<td>10</td>
<td>7620</td>
<td>7781</td>
<td>90</td>
<td>9.2</td>
<td>9.9</td>
<td>Portal protein</td>
<td>ORF339 phage O1205 (108/112; 96%)</td>
</tr>
<tr>
<td>11</td>
<td>7793</td>
<td>8136</td>
<td>105</td>
<td>11.3</td>
<td>10.3</td>
<td>Capsid protein</td>
<td>ORF340 phage O1205 (108/112; 96%)</td>
</tr>
<tr>
<td>12</td>
<td>8131</td>
<td>8447</td>
<td>116</td>
<td>12.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF341 phage O1205 (108/112; 96%)</td>
</tr>
<tr>
<td>13</td>
<td>8448</td>
<td>8785</td>
<td>336</td>
<td>38.7</td>
<td>9.9</td>
<td>DNA binding</td>
<td>S. pyogenes prophage 315.5 (24/40; 60%)</td>
</tr>
<tr>
<td>14</td>
<td>8787</td>
<td>9173</td>
<td>387</td>
<td>41.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF342 phage Sfi11 (110/119; 92%)</td>
</tr>
<tr>
<td>15</td>
<td>9178</td>
<td>9565</td>
<td>501</td>
<td>57.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF343 phage O1205 (472/501; 94%)</td>
</tr>
<tr>
<td>16</td>
<td>9772</td>
<td>10125</td>
<td>353</td>
<td>37.3</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF344 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>17</td>
<td>10176</td>
<td>10493</td>
<td>318</td>
<td>35.0</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF345 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>18</td>
<td>10493</td>
<td>15036</td>
<td>455</td>
<td>48.7</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF346 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>19</td>
<td>15036</td>
<td>16541</td>
<td>105</td>
<td>12.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF347 phage Sfi11 (110/119; 92%)</td>
</tr>
<tr>
<td>20</td>
<td>16571</td>
<td>21388</td>
<td>1605</td>
<td>177.3</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF348 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>21</td>
<td>21389</td>
<td>25010</td>
<td>3626</td>
<td>39.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF349 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>22</td>
<td>25011</td>
<td>30904</td>
<td>5893</td>
<td>55.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF350 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>23</td>
<td>30905</td>
<td>32609</td>
<td>1704</td>
<td>17.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF351 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>24</td>
<td>32609</td>
<td>33654</td>
<td>1045</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF352 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>25</td>
<td>33655</td>
<td>34702</td>
<td>1047</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF353 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>26</td>
<td>34703</td>
<td>36454</td>
<td>1752</td>
<td>19.6</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF354 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>27</td>
<td>36455</td>
<td>37500</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF355 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>28</td>
<td>37501</td>
<td>38547</td>
<td>1047</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF356 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>29</td>
<td>38548</td>
<td>39593</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF357 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>30</td>
<td>39594</td>
<td>40639</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF358 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>31</td>
<td>40640</td>
<td>41695</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF359 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>32</td>
<td>41696</td>
<td>42741</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF360 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>33</td>
<td>42742</td>
<td>43787</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF361 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>34</td>
<td>43788</td>
<td>44833</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF362 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>35</td>
<td>44834</td>
<td>45879</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF363 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>36</td>
<td>45880</td>
<td>46925</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF364 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>37</td>
<td>46926</td>
<td>47971</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF365 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>38</td>
<td>47972</td>
<td>48917</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF366 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>39</td>
<td>48918</td>
<td>49963</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF367 phage Sfi11 (326/348; 93%)</td>
</tr>
<tr>
<td>40</td>
<td>49964</td>
<td>50069</td>
<td>1046</td>
<td>11.4</td>
<td>9.9</td>
<td>Capsid protein</td>
<td>ORF368 phage Sfi11 (326/348; 93%)</td>
</tr>
</tbody>
</table>

<sup>a</sup>—, unknown function; TMP, tape measure protein; SSAP, single-strand annealing protein.
FIG. 3. Alignment of the genetic maps of all completely sequenced phage genomes. The modular regions of the genomes coding for distinct functions are indicated above the maps. Deduced proteins sharing more than 50% amino acid identity are represented using the same color and are linked using grey shading whenever possible. ORFs with unique sequences are displayed in white. Genes coding for proteins identified by N-terminal sequencing or MALDI-TOF are identified by thick lines.
prophage in a *S. thermophilus* host, as demonstrated for other LAB phages (6, 24, 51). Alternatively, phage 2972 may have started out as a temperate phage that became virulent following deletion(s) and/or rearrangement(s) leading to a nonfunctional lysogeny module. It has been demonstrated that lytic phages can emerge after several passages of the temperate *S. thermophilus* phage Sfi21 on an indicator strain (11, 13).

**Morphogenesis.** Comparative genomic analysis has clearly demonstrated the presence of two clusters of morphogenesis genes in *S. thermophilus* phages (9, 20). These two clusters support the existence of two *S. thermophilus* phage groups, the cos and pac types. The morphogenesis genes of phage 2972 are in line with this grouping, as they are homologous with the morphogenesis genes of the pac-type phages Sfi11 and O1205 (Fig. 3). Because the morphogenesis module of *S. thermophilus* phages has already been extensively described elsewhere (9, 20), we will focus here on the divergences and novelties observed in phage 2972.

The putative tail protein ORF19 is one of the most intriguing gene products of the deduced proteome of phage 2972. As shown in Fig. 4, many conserved amino acids have been found in other *S. thermophilus* phage proteins that are similar to ORF19 of phage 2972. In general, the N- and C-terminal portions of these proteins are similar among members of the cos- and pac-type groups. However, the central region of proteins similar to ORF19 of phage 2972 (approximately aa 225 to aa 410) is conserved in both groups. ORF19 of phage 2972 shares 62% amino acid identity (282/449) with ORF17 of the cos-type phage DT1 and 57% identity (298/518) with ORF512 of the pac-type phage Sfi11 (Table 1). Overall, ORF19 of phage 2972 may thus be a hybrid structural protein that connects the two *S. thermophilus* phage groups.

Comparison of ORF20 with the deduced proteome of the other pac-type phages revealed that it was the most divergent structural protein. The function of ORF20 may be to recognize the specific phage receptor on the streptococcal surface. Indeed, ORF20 exhibited some degree of identity with the receptor-binding protein of the *S. thermophilus* cos-type phage MD2 (23). Receptor-binding proteins from phages infecting low-G+C-content gram-positive bacteria usually contain collagen-like repeat motifs at their C termini (60). This motif appears to be characteristic of collagen molecules, and its biological function is to provide elasticity and confer stability on the triple helix structure (4). Six collagen-like repeats were found in ORF20 of phage 2972 (Fig. 5). Three variable regions are found in ORF20, and two of them (VR1 and VR3) are flanked by collagen-like repeats as observed in other streptococcal phages (23). Interestingly, VR1 of ORF20 is also present in ORF38 of the cos-type phage 7201 and VR3 is also present in ORF45 of the pac-type phage O1205 (Fig. 5). These observations illustrate the modular organization of the putative receptor-binding proteins. As also shown in Fig. 5, the putative receptor-binding proteins are unique to each cos- and pac-type phage, except for the conserved C-terminal regions of
the deduced proteins. These findings are consistent with the fact that phages O1205, Sfi11, Sfi19, DT1, and 7201 cannot propagate on the host strain of phage 2972 (data not shown).

Protein composition of phage 2972. CsCl-purified phage particles were analyzed by SDS-PAGE in order to identify the proteins in the virion structure (Fig. 6). Eight structural proteins were identified by N-terminal sequencing and/or MALDI-TOF. They included the putative portal protein (ORF5), the major capsid protein (ORF9), the major tail proteins (ORF15 and ORF17), three putative minor tail proteins (ORF18, ORF19, and ORF21), and the putative receptor-binding protein (ORF20).

The identification of ORF21 in the structure of phage 2972 is interesting, as it indicates that this gene is part of the morphogenesis module. Previous genomic analyses were inconclusive in predicting its classification as a structural or a nonstructural protein (20). According to the comparative analysis presented in Table 1 and Fig. 3, this structural protein appears to be relatively conserved in *S. thermophilus* cos- and pac-type phages. For example, ORF21 of phage 2972 shares 52% identity (345/663 amino acids) with ORF19 of the cos-type phage.

---

**FIG. 5.** Schematic illustration representing the alignment of the proteins possibly involved in host recognition of seven *S. thermophilus* phages. The same color indicates more than 80% similarity. Collagen-like repeats are shown in red. VR1, VR2, and VR3 indicate variable regions 1, 2, and 3, respectively. Regions with unique sequences are shown in white.

**FIG. 6.** Protein profiles of phages 2972 (pac type) and DT1 (cos type) as determined on SDS-PAGE gels stained with Coomassie blue. MW, molecular weight markers.
DT1 (Table 1). This is in contrast with the current view that these two groups of phages have different sets of morphogenesis genes. To confirm that ORF19 of phage DT1 is also present in the virion structure, CsCl-purified phage particles were analyzed by SDS-PAGE (Fig. 6). By using N-terminal sequencing, a protein of ~76 to 78 kDa was identified as ORF19, confirming that this protein is indeed present in the structure of this *S. thermophilus* cos-type phage.

**Introns in the phage 2972 genome.** As indicated above, sequence analysis suggested the presence of two introns in the genome of phage 2972. The first intron has been located between *orf3* and *orf4* within the gene coding for the putative terminase large subunit (*terL*-1), while the second intron has been found between *orf26* and *orf29*, interrupting the endolysin-encoding gene (*lys*-1). An intron interrupting the endolysin-encoding gene has already been characterized in other *S. thermophilus* phages (25), and a group I intron that interrupts the gene encoding the large subunit of the terminase of the virulent phage LL-H of *Lactobacillus delbrueckii* has also been reported (47).

To test for in vivo splicing of RNA transcripts, RT-PCR experiments were performed. By using specific primers located in *orf26* and *orf29* of phage 2972, a PCR product of 590 bp was amplified from the phage genomic DNA, while a 148-bp amplicon was obtained using the cDNA as a template (data not shown). Sequence analysis of the PCR products revealed that the splicing occurred after a uridine residue (coordinate 25280 within *orf26*) as well as after a guanosine residue (coordinate 25724 upstream of *orf29*), resulting in the excision of a 442-bp intron. The endolysin-encoding gene is thus 843 bp long, and it is believed to code for a 281-aa protein that possesses 79% identity (219/275) with the endolysin of *S. thermophilus* phage S3b. The splicing occurred at exactly the same site as that observed for the intron of phage S3b (25). The secondary structure of this intron was relatively similar to that of phage S3b (25), except that the P3.1 and P3.2 stems were included in the prediction (Fig. 7A). The P7.2 stem folding retained was similar to that of phage SP01. The main nucleotide discrepancies were located within the P8 looped-out region. One other notable difference was in the P7.2 stem, where a guanine was present at coordinate 25438, compared to an adenine in phage SB3. An additional adenosine was also found in the looped-out region of P3.2, creating a short ORF (*orf27*, coding for 40 amino acids).

For the second intron, the in vivo splicing of the mRNA was demonstrated using specific primers located in *orf3* and *orf4*. A 506-bp DNA fragment was amplified from the phage genomic DNA, while a 199-bp PCR product was obtained using cDNA as a template (Fig. 8A). Sequence analysis of the PCR products revealed that the splicing occurred after a uridine and a guanosine residue (coordinates 2098 and 2406, respectively). The 307-bp intron had a G+C content similar to the rest of the phage genome, did not contain an ORF, and had 83% identity (55/60) with a group IAI intron found in the chloroplastic gene coding for the rRNA large subunit of a *Chlamydomonas* sp. (GenBank accession no. L43539). After the mRNA splicing, the *terL* gene of phage 2972 was 1,236 bp long and coded for a 411-aa protein (Fig. 8B) that possessed 96% identity (394/410) with the intron-free ORF411 of *S. thermophilus* phage Sfi11 and 95% identity (393/410) with ORF26 of phage O1205.

Analysis of the DNA region flanking the integration site of this intron in phage 2972 with the corresponding region in phages Sfi11 and O1205 revealed sequence variations close to the integration site that may explain the absence of an intron in these two phages (Fig. 8C). The secondary-structure prediction (15, 18, 46) of the terminase intron possessed all the canonical group I intron features (P1 to P9) that are required to form the catalytic core of the intron and that are essential for self-splicing activity (Fig. 7B). The P7.1 and P7.1a stem-loops between P3 and P7 are characteristics of a subgroup IA1 intron. An internal guide sequence that could bring P1 and P10 into close proximity to facilitate the splicing process (16, 17) was also recognized.

Lastly, structural differences were noted between the two introns found in the genome of phage 2972. *terL*-1 possessed the P2, P5a, and P9.1 stems, while *lys*-1 had the P3.1, P3.2, and P7.2 stems as well as an ORF (Fig. 7). The two introns thus belong to different subgroups (IA1/terminase, IA2/endolysin) and are probably from different sources.

**DISCUSSION**

**Phage sensitivity and exopolysaccharide production.** We demonstrated here that several phages can infect an EPS-producing *S. thermophilus* strain such as RD534. Some studies have previously suggested that cell surface phage receptors can be blocked by the loosely bound EPS produced by some bacterial strains, thus protecting the cells against phage infections (7, 29, 30, 65). Brissow et al. (10) previously reported the isolation of phages (including Sfi11, Sfi19, and Sfi21) from ropy strains of *S. thermophilus*. We also recently reported that the EPS-producing *S. thermophilus* strain MR-1C and its EPS-negative derivative were both sensitive to the same three pac-type phages (7). Several phages can also infect EPS-producing *Lactococcus lactis* strains (21). Taken together, these results clearly indicate that the production of EPS does not confer potent protection against phage infections.

**Genome of phage 2972.** We presented the seventh complete genome of an *S. thermophilus* phage, and the third from the pac-type group. With its 34,704 bp, phage 2972 possesses the shortest *S. thermophilus* phage genome analyzed so far. As previously reported, the lysogeny region may be a recombination hot spot in *S. thermophilus* phages (40). The presence of a *cro*-like repressor gene suggests that virulent phage 2972 is derived from a temperate phage. It is noteworthy that a *cro*-like repressor gene is present in many virulent *S. thermophilus* phage genomes (40), and one might wonder whether the *cro*-like repressor still plays a role in the lytic cycle of virulent phages, particularly in lysogenic hosts.

Another noteworthy size variation among the three pac-type phages was noted in the genome area coding for the tail proteins. Because of its position in the genome and its similarity to ORF18 of phage DT1, which was experimentally shown to be involved in the recognition of *S. thermophilus* hosts (23), ORF20 is most likely involved in host recognition.

**Introns of phage 2972.** Two group I introns have been found in the genome of phage 2972. They were located in the genes coding for the terminase large subunit and the endolysin, where introns have been found in other phage genomes (25, 47). Phage introns seem to target crucial genes in the phage
FIG. 7. Secondary-structure predictions of the two introns in the genome of phage 2972. The secondary-structure representation was made using a two-dimensional structural diagram (15, 18, 46). Arrows indicate the 5' and 3' splice sites. Lower- and uppercase letters denote the exon and intron sequences, respectively. Boxed sequences indicate the regions that anneal to form P9.0 and P10. The shaded nucleotides in P7 represent the putative guanosine-binding site. Bold lines show connections between intron sequence domains, with pointers indicating the 5'-to-3' direction. IGS, internal guide sequence. Numbers in parentheses represent the nucleotide position on the phage 2972 genome.

(A) Intron in the gene coding for the endolysin. The start and stop codons of ORF28 are underlined, and the two nucleotides that differ from the S3b intron are boldfaced. (B) Intron in the gene coding for the large subunit of the terminase.

VOL. 71, 2005
GENOMIC SEQUENCE OF
S. THERMOPHILUS
PHAGE 2972
4065
phage introns of the IA1 subgroup to be characterized. Recently, a subgroup-IA1 intron was uncovered in the genome of the Synechococcus cyanophage S-PM-2 (48). It interrupted a gene (psbA) coding for a core component of the photosynthetic reaction center PSII (photosystem II) (48). Group I introns range in size from 200 to 3,000 bp, depending on the length of the peripheral sequence and on whether or not they contain ORFs (33). The terL-I is a relatively short, 307-bp intron that does not contain an ORF, in contrast to phage Twort orf142-11, -12, and -13 (34) and phage S-PM2 psdA-I (48).

**Structural phage proteins and phage classification.** The identification of a conserved structural protein (ORF21) in both groups of *S. thermophilus* phages (cos and pac types) is interesting considering the fact that these two groups are regarded as two lineages of the family Siphoviridae (54). Moreover, the discovery of the structural hybrid protein ORF19 in phage 2972 suggests that recombination may have occurred within these two distinct structural gene clusters, possibly during mixed infections. While the functions of ORF19 and ORF21 remain to be determined, the position of their genes in the genome suggests that they are tail-related proteins. orf20 possibly codes for the receptor-binding protein of phage 2972. It is not known whether these three proteins interact with each other, but it appears that this area of the genome has the flexibility and potential for domain shuffling. Such rearrangements may favor the formation of functional recombinant phages with a modified host range.

Despite the finding of a common structural protein, the current classification of *S. thermophilus* phages based on DNA packaging mechanisms (cos and pac) and structural protein composition remains valid. *S. thermophilus* phages were first classified into a single DNA homology group based on DNA-DNA hybridization data, a classification supported by their similar morphology (45). Since then, comparative analyses of a growing number of streptococcal phage sequences have confirmed these conserved genomic regions. The presence of hybrid and conserved structural proteins provides new evidence to support the hypothesis of a common ancestor. The subsequent sorting of these phages into two phage groups is also quite evident based on the overall makeup of their genomes and proteomes. A view toward practical applications is perhaps the most compelling reason for maintaining the current classification of *S. thermophilus* phages in two groups. In our experience, most phage-sensitive *S. thermophilus* strains are infected either by cos-type phages or by pac-type phages. A given phage-sensitive *S. thermophilus* strain is rarely infected by members of both phage groups. For example, *S. thermophilus* RD534 has been infected only by pac-type phages, such as 2972. Consequently, it has been possible to design or rotate starter cultures based on group sensitivity.

On a larger scale, a number of proposals have been put forward recently to modify the current approach of the International Committee on Taxonomy of Viruses (35, 54, 55). The study reported here supplies evidence that proposals based solely on structural gene modules may not be the answer. Lastly, it should be remembered that such proposals not only need to be scientifically sound but should also be useful from an applied perspective.

---

**FIG. 8. Characterization of the intron between the genes coding for the terminase large subunit of phage 2972.** (A) In vivo splicing of 2972 intron RNA (terminase large-subunit gene). Lane 1, 100-bp marker (Invitrogen); lane 2, negative control without DNA, cDNA, or RNA; lane 3, PCR product obtained with 2972 DNA as a template; lane 4, PCR product obtained with RNA isolated from 2972-infected *S. thermophilus* cells (cDNA); lane 5, PCR product (no reverse transcription) obtained with RNA isolated from 2972-infected cells. (B) Phage 2972 genomic region containing the intron and the two ORFs coding for the terminase large subunit. Large, thick arrows indicate open reading frames; the numbers of amino acids (aa) in the deduced proteins are given. Splicing of the 307-bp intron resulted in a 411-aa protein. Small arrows represent the primers used for PCR and sequencing. (C) Nucleotide sequence alignment of the regions flanking the splicing site with the corresponding intron-free regions in two other *S. thermophilus* phages (Sfi11 [orf411] and O1205 [orf26]). Differences relative to the phage 2972 sequence are indicated. Vertical arrow indicates the intron insertion site. The nucleotide positions are based on the 2972 genomic sequence (GenBank accession no. AY099705).

---

genome (DNA polymerase, thymidylate synthase, ribonucleotide reductase, structural proteins, terminase large subunit, and endolysin), while introns in eubacteria are in tRNA genes and all belong to a different subgroup (IC3) (25, 36). The distribution and the homing of group I introns have been studied for T-even-like bacteriophages, and the results suggested that these introns share a recent common ancestor that is not known whether these three proteins interact with each other, but it appears that this area of the genome has the flexibility and potential for domain shuffling. Such rearrangements may favor the formation of functional recombinant phages with a modified host range.

Despite the finding of a common structural protein, the current classification of *S. thermophilus* phages based on DNA packaging mechanisms (cos and pac) and structural protein composition remains valid. *S. thermophilus* phages were first classified into a single DNA homology group based on DNA-DNA hybridization data, a classification supported by their similar morphology (45). Since then, comparative analyses of a growing number of streptococcal phage sequences have confirmed these conserved genomic regions. The presence of hybrid and conserved structural proteins provides new evidence to support the hypothesis of a common ancestor. The subsequent sorting of these phages into two phage groups is also quite evident based on the overall makeup of their genomes and proteomes. A view toward practical applications is perhaps the most compelling reason for maintaining the current classification of *S. thermophilus* phages in two groups. In our experience, most phage-sensitive *S. thermophilus* strains are infected either by cos-type phages or by pac-type phages. A given phage-sensitive *S. thermophilus* strain is rarely infected by members of both phage groups. For example, *S. thermophilus* RD534 has been infected only by pac-type phages, such as 2972. Consequently, it has been possible to design or rotate starter cultures based on group sensitivity.

On a larger scale, a number of proposals have been put forward recently to modify the current approach of the International Committee on Taxonomy of Viruses (35, 54, 55). The study reported here supplies evidence that proposals based solely on structural gene modules may not be the answer. Lastly, it should be remembered that such proposals not only need to be scientifically sound but should also be useful from an applied perspective.
ACKNOWLEDGMENTS

We thank Diane Montpetit for assistance with electron microscopy, Jean-François Pombert and Christian Ors for assistance in secondary-structure analysis of the introns, Dennis Romero for valuable discussions, and Gene Bourgeois for editorial assistance.

This work was funded, in part, by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES


