**Cell surface glycoproteins from Thermoplasma acidophilum are modified with an N-linked glycan containing 6-C-sulfofucose†**

Evgeny Vinogradov, Lise Deschatelets, Marc Lamoureux, Girishchandra B Patel, Tammy-Lynn Tremblay, Anna Robotham, Marie-France Goneau, Cathy Cummings-Lorbetskie, David C Watson, Jean-Robert Brisson, John F Kelly, and Michel Gilbert

Human Health Therapeutics, National Research Council Canada, Ottawa, Ontario, Canada K1A 0R6

Received on March 3, 2012; revised on May 31, 2012; accepted on June 2, 2012

_Thermoplasma acidophilum_ is a thermoacidiphilic archaeon that grows optimally at pH 2 and 59°C. This extremophile is remarkable by the absence of a cell wall or an S-layer. Treating the cells with Triton X-100 at pH 3 allowed the extraction of all of the cell surface glycoproteins while keeping cells intact. The extracted glycoproteins were partially purified by cation-exchange chromatography, and we identified five glycoproteins by N-terminal sequencing and mass spectrometry of in-gel tryptic digests. These glycoproteins are positive for periodic acid-Schiff staining, have a high content of Asn including a large number in the Asn-X-Ser/Thr sequon and have apparent masses that are 34–48% larger than the masses deduced from their amino acid sequences. The pooled glycopeptides were digested with proteinase K and the purified glycopeptides were analyzed by NMR. Structural determination showed that the carbohydrate part was represented by two structures in nearly equal amounts, differing by the presence of one terminal mannose residue. The larger glycan chain consists of eight residues: six hexoses, one heptose and one sugar with an unusual residue mass of 226 Da which was identified as 6-deoxy-6-C-sulfo-D-galactose (6-C-sulfofucose). Mass spectrometry analyses of the peptides obtained by trypsin and chymotrypsin digestion confirmed the principal structures to be those determined by NMR and identified 14 glycopeptides derived from the main glycoprotein, Ta0280, all containing the Asn-X-Ser/Thr sequons. _Thermoplasma acidophilum_ appears to have a “general” protein N-glycosylation system that targets a number of cell surface proteins.

**Keywords:** Thermoplasma / glycoprotein / N-linked glycan / sulfofucose

**Introduction**

_Thermoplasma acidophilum_ is a thermoacidiphilic archaeon that was first isolated from a coal refuse pile that had undergone self-heating (Darland et al. 1970). It grows optimally at pH 2 and 59°C but is able to multiply between pH 0.5 and 4 and at temperatures between 40 and 62°C. The cells are spherical and variable in size, ranging from 0.1 to 5 μm. Although _T. acidophilum_ thrives in a harsh environment, its cells lack a cell wall or a surface (S) layer and are delineated by a single double-layer membrane. These characteristics have prompted research on the structural basis for the stability of the _T. acidophilum_ membrane under extreme environmental conditions. The membrane contains ether lipids based on 40-carbon, isoprenoid-branched diglycerol tetraethers (Langworthy et al. 1972). Yasuda et al. (1995) reported fibrous structures on the surface of _T. acidophilum_ using transmission electron microscopy to observe thin sections. Ruwart and Haug (1975) reported that detergents caused the partial solubilization of cellular material while leaving the cells intact.

Post-translational glycosylation of proteins is now well documented in both Bacteria and Archaea (Abu-Qarn et al. 2008; Yurist-Doutsch et al. 2008; Nothaft and Szymanski 2010). However, detailed structural information is available only for a limited number of prokaryotic glycans. N- and O-linked glycans have been reported for S-layer glycoproteins from Archaea (Calo et al. 2010; Jarrell et al. 2010), whereas _O_-linked glycans have been observed for _S_-layer glycoproteins from Bacteria (Messner et al. 2008; Posch et al. 2011; Ristl et al. 2011). There are also multiple examples of prokaryotic glycosylated proteins other than _S_-layer proteins (Abu-Qarn et al. 2008; Nothaft and Szymanski 2010).
Glycosylation of flagellins has been reported for both Archaea (Jarrell et al. 2010) and Bacteria (Nothaft and Szymanski 2010). A “general” N-glycosylation pathway that transfers a heptasaccharide to periplasmic proteins has been well characterized in Campylobacter jejuni (Szymanski et al. 1999; Wacker et al. 2002; Young et al. 2002).

Yang and Haug (1979) reported the isolation of a glycoprotein with a mass of 152 kDa from T. acidophilum membranes and performed preliminary determination of the glycan structure. These data suggested that the 152-kDa protein is a surface glycoprotein with a glycan portion similar to eukaryotic high-mannose N-linked glycans (Yang and Haug 1979). In this work, we report the identification of this major glycoprotein and other cell surface glycoproteins, as well as the complete structural determination of the N-linked glycan.

Results
General detection of glycoproteins and their cell localization
We used a high-pressure homogenizer to disrupt the cells in Tris buffer, pH 7.5, and differential centrifugation to separate the cell extract in three fractions: cell debris (12,000 × g pellet), membranes (260,000 × g pellet) and soluble proteins (260,000 × g supernatant). Using sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS–PAGE) and periodic acid-Schiff (PAS) staining, we detected glycosylated material in the cell debris and membrane fractions but not in the soluble protein fraction (data not shown). This observation was consistent with Yang and Haug (1979) who reported that two glycoproteins were associated with membranes in T. acidophilum. While evaluating various conditions for extracting the glycoproteins, we observed that all of the glycosylated material could be extracted directly from intact cells using 1% Triton X-100 in citrate buffer, pH 3 (Figure 1). The protocol first involved washing the cells with citrate buffer (pH 3) to remove loosely bound proteins (Figure 1A, lane 3) that appeared not to be glycosylated (Figure 1B, lane 3). After treating the cells with 1% Triton X-100 in citrate buffer, pH 3, the bulk of the proteins remained associated with the cells (Figure 1A, lane 5), whereas all of the glycosylated material (detected by PAS staining) was recovered in the supernatant after centrifugation at 12,000 × g (Figure 1A and B, lane 4). We extracted 8–10 mg of glycoproteins from 1 g of cell paste (wt weight). Four faint bands were detected by both Coomassie and PAS staining after SDS–PAGE of the Triton extract (Figure 1A and B, lane 4), suggesting that at least four glycoproteins were extracted from the cell surface. We determined the apparent molecular masses of the four main glycoprotein bands to be 60, 76, 120 and 150 kDa, respectively. The PAS staining also revealed streaking material that ran across the whole lane (Figure 1B, lane 4), suggesting that there is also glycosylated material that is not conjugated to proteins, since this streaking material was not stained by Coomassie blue (Figure 1A, lane 4). The streaking material detected by PAS staining probably corresponded to the lipoglycan reported by Mayberry-Carson et al. (1974) (see Discussion).

Partial purification and identifications of the major glycoproteins
The proteins extracted with citrate/Triton X-100 were partially purified by cation-exchange chromatography in sodium acetate, pH 4, with 0.1% Triton X-100. Most of the glycoproteins bound to the column under these conditions (Figure 2, lanes 6–8), whereas the streaking material that stained with PAS did not bind and was recovered in the flow through (Figure 2B, lane 4). The glycoproteins eluted between 0.3 and 1 M NaCl with low resolution between the main glycoproteins.

We used two procedures to identify the main glycoproteins: N-terminal sequencing of proteins transferred to a
polyvinylidene fluoride (PVDF) membrane and mass spectrometry of tryptic in-gel digests (Table I). N-terminal sequencing identified four protein sequences: Ta0280 for band #1 (150 kDa), Ta1063 for band #3 (76 kDa), Ta0171 for band #5 (48 kDa) and Ta0171 for band #5 (48 kDa). An N-terminal sequence could not be determined for band #2 (120 kDa) which could be due to a blocked N terminus or to weak signals resulting from a heterogeneous N-terminal sequence. It is notable that the apparent masses of the proteins as measured by SDS–PAGE on a 8% gel are 34–48% larger than the masses deduced from their amino acid sequences (Table I). Furthermore, each protein sequence possesses an unusually large number of N-linkage sequons (Table I). Taken together, the evidence suggests that these are all good candidates to be glycoproteins with N-linked glycans.

Mass spectrometry analysis of the tryptic in-gel digests identified the same proteins as the ones identified by N-terminal sequencing in the cases of bands #1, #4 and #5 (Table I and Supplementary data, Table S1). Three proteins were identified in band #2 (120 kDa), one of which, Ta0261, is unique to this band whereas the other two are putative glycoproteins that were already identified in bands #1 and #4 (Ta0280 and Ta1052, respectively). Ta0261 is a good candidate as a glycoprotein, having a high content of N-linkage sequons and a deduced mass (82.3 kDa) that is significantly smaller than the mass determined by SDS–PAGE for band #2 (120 kDa). Eight proteins were identified in the digest of band #3 (Supplementary data, Table S1), including Ta1063, the only sequence seen by N-terminal sequencing of this band. N-terminal sequencing is significantly less sensitive than mass spectrometry and it is possible that Ta1063 is actually the main protein component in band #3. We propose that Ta1063 is the most likely candidate to be the 76-kDa glycoprotein as it is the only one of the eight identified by mass spectrometry to possess a large number of N-linkage sequons (Supplementary data, Table S1) and a determined mass (76 kDa) that is significantly larger than the deduced mass (55.5 kDa).

The open reading frames corresponding to the five identified glycoproteins are all annotated as hypothetical proteins in the genome sequence of T. acidophilum DSM 1728 (GenBank accession number NC_002578) that was determined by Ruepp et al. (2000). PSI-BLAST searches (NCBI) performed with the sequences of the five glycoproteins returned alignments with various hypothetical proteins from other archaeons (data not shown). Ta0280, the major glycoprotein, shares 65% with a hypothetical protein from Thermoplasma volcanium (a close species) and 30% identity with a hypothetical protein from Picrophilus torridus, another thermoacidophilic archean. Ta1063 also shows significant homology only with hypothetical proteins from other thermoacidophilic archaeons, but a putative conserved domain (COG1361, S-layer domain) was detected in this glycoprotein. Putative conserved domains were also detected in Ta0171 (CD01143, initial receptors in ATP-binding cassette transport of metal ions), Ta0261 (CD08512, substrate-binding domain of ATP-binding cassette transporters) and Ta1052 (COG0725, ABC-type inorganic ion transport system). PSI-BLAST searches with these three glycoproteins returned alignments with various archael hypothetical proteins, including some that are annotated as ATP-binding cassette transporters or solute-binding proteins which are consistent with the conserved domains that were detected in their respective sequences.

Analysis of the protein sequences using PRED-SIGNAL and SignalPv3 suggests that all five glycoproteins have a signal peptide for secretion. Although Ta0171 (48 kDa) and Ta0261 (120 kDa) were detected in the cell growth medium at very low concentration (data not shown), the bulk of the glycoproteins needed to be extracted from cells with Triton X-100 which indicated that they mostly remained associated with the cells even if they contained a predicted secretion signal peptide. Analyses of the protein sequences for transmembrane domains with dense alignment surface (DAS) predicted a C-terminal transmembrane domain for Ta0280 and Ta1063, and an N-terminal transmembrane domain for Ta0171, Ta0261 and Ta1052. The C-terminal transmembrane domains of Ta0280 and Ta1063 would explain why these glycoproteins would remain associated with the cells following their translocation. Based on their N-terminal sequences, Ta0171 and Ta1052 have their N-terminal transmembrane domains intact which would account for their association with the cells. Since we did not obtain an N-terminal sequence for Ta0261, we cannot determine if its N-terminal transmembrane domain can account for its association with the cells.

**NMR analyses of the glycopeptides obtained by proteinase K digestion**

Since we could not isolate individual glycoproteins, we designed a strategy to determine the structure of the N-linked glycan from a pool of glycoproteins extracted from intact cells and partially purified on a cation-exchange column. We combined the fractions that eluted between 0.2 and 1 M
NaCl which contained the main glycoproteins. These pooled glycoproteins were digested with proteinase K and the glycopeptides were purified by gel chromatography and anion-exchange chromatography (see Materials and methods). NMR data (Figures 3 and 4) showed that the carbohydrate part was represented by two structures 1a and 1b in nearly equal amounts, differing by the presence of one terminal mannose residue (Figure 5). The structure was analyzed using selective one-dimensional NMR [total correlation spectroscopy (TOCSY) and nuclear Overhauser effect spectroscopy (NOESY)] and two-dimensional NMR [NOESY, TOCSY, double quantum filtered correlation spectroscopy, $^1$H-$^1$C heteronuclear single quantum coherence (HSQC) and heteronuclear multiple bond coherence (HMBC)] experiments. Spin systems of 10 monosaccharides were identified; however, some of them actually belonged to the same monosaccharides (B and C, D and E) from compounds 1a (octasaccharide) or 1b (heptasaccharide). Heterogeneity due to 50% of residue H being present leads to sequences H-B-D and C-E being observed. Identity of the monosaccharides was deduced from the coupling constants and signal positions obtained from one-dimensional TOCSY and NOESY experiments (Supplementary data, Figures S1 and S2).

Two monosaccharides required a special approach for their identification. $\alpha$-glycero-$\alpha$-o-galacto-heptose (A) had the obvious $\alpha$-galacto-configuration of the pyranose ring from the measurement of the proton coupling constants. Configuration of C-6 was identified by the comparison of the NMR data with the synthesized model monosaccharide. The chemical shift of C-6 near 70 ppm was also characteristic for $\alpha$-glycero C-6 (Vinogradov et al. 2004). The 6-C-sulfo-$\alpha$-fucose (6-deoxy-6-C-sulfo-$\alpha$-galactose, Fuc6S) residue had an $\alpha$-galactopyranose ring and the specific position of its H/C-6 signals at 3.29–3.25/52.8 ppm agreed with the values previously described for this monosaccharide (Fernández-Bolaños et al. 2003). The absolute $\alpha$-configuration of all Glc, Gal and Man residues was postulated as found in nature, and the absolute configuration of the Hep and Fuc6S was determined by nuclear Overhauser effects (NOEs) and molecular modeling.

The Asn-linked $\beta$-Gal (F) was identified on the basis of a high-field position of its C-1 signal at 80.3 ppm, which is characteristic for $N$-linked $\beta$-sugars. The large intra-ring NOEs F1:F3 and F1:F5 confirmed the $\beta$-configuration of the pyranose ring (Figure 3B). Intra-ring NOEs from H-1 to H-3,5 resonances were also observed for two other $\beta$-sugars, residues I and J (Figure 3C).

The sequence of the monosaccharides was determined from NOE (Figure 3) and HMBC (Figure 4) data. Thus, the strong inter-residues NOE J1:I4 establish the J-4I-3F sequence (Figure 3C). The A1:J2,J3 NOE was due to the A-2J sequence and the strong coupling between J2 and J3 (Figure 3D). NOEs DE1:J2,J3,J4 were all due to the linkage DE-3J (Figure 3E). The determination of the arrangement of substituents A and D,E on the Fuc6S residue presented some difficulty: H-2 and H-3 of Fuc6S nearly exactly overlapped; HMBC correlations from A1 were doubtful due to the overlap of carbon signals A3, A5 and J2. The NOE from DE1 to J4 was very strong, which is not unusual but still required additional analysis. Only the connection of D,E to J3 was confirmed by HMBC (Figure 4). To finally resolve this part of the structure, oligosaccharides were acetylated. This caused protons at acetylation points to move far to the low field and remove the signal overlap. The NMR data for the residues A, D,E and J in acetylated 1a,b (Table II; NOE not shown).

**Fig. 3.** One-dimensional Proton NOE spectra of pooled glycopeptides. (A) $^1$H spectrum; (B) NOESY1D from F1; (C) NOESY1D from I1 and J1; (D) NOESY1D from A1; (E) NOESY1D from D1,E1; (F) NOESY1D from B1,C1; (G) TOCSY-NOESY1D from D6,E6,H2 at 4.02 ppm in the first TOCSY step with a mixing time of 20 ms to locate DE6, which was then selected for the NOESY step. The inter-residue DE6-G1 NOE was observed establishing the G-6DE sequence. (H) NOESY1D from H1 shows the H-6B sequence. A mixing time of 800 ms was used for the NOESY1D experiments.
clearly proved that the presented structure was correct. The Man(1-3)Man sequence was obtained from the B1-D3 and C1-E3 NOE (Figure 3F) and the B1:D3 and C1:E3 HMBC correlation (Figure 4). The presence of residue H (50%) linked to Man(1-3) leads to different chemical shifts for these Man(1-3)Man sequences in 1a and 1b.

Molecular modeling
Long-range NOEs between residues that are not directly linked were used to determine the absolute configuration of the Hep (A) and Fuc6S (J). These NOEs included the A1:J5 NOE (Figure 3D) in the sequence A-J-I and the B1C1:A4 NOE (Figure 3F) in the sequence BC-DE-J-A which was also used to confirm the 2,3 branching point on residue J. The D1E1-J4 NOE (Figure 3E) for Man(1-3)Fuc6S was also dependent on the absolute configuration of Fuc6S. With the absolute α-configuration postulated for the Gal, Glc and Man residues as found in nature, the observed NOEs could only be observed if Hep and Fuc6S have the absolute D-configuration (Supplementary data, Figure S3).

Mass spectrometry analysis of the glycan
Liquid chromatography-mass spectrometry (LC-MS) analysis of the in-gel tryptic digest of Ta0280 detected numerous glycopeptides with glycoform profiles similar to that shown in Figure 6A for the tryptic peptide, T122–139. The two predominant glycoforms differ from one another by a single hexose sugar and correspond in mass to the 1a and 1b structures determined by NMR (Hex5-6Hep1dHexS1). The mass spectrometry/mass spectrometry (MS/MS) spectra were dominated by ions arising from the fragmentation of the glycan moieties but often contained weak peptide fragment ions that confirmed their amino acid sequences (Figure 6B). The pattern of sugar losses as revealed by the doubly charged fragment ion series at m/z 1000–1700 in the MS/MS spectrum presented in Figure 6B indicates that the glycan is linked to the peptide via two hexose moieties and that the 3rd and 4th sugars are the Fuc6S (226 Da) and heptose (192 Da) moieties, respectively.
with the remainder of the glycan being composed of hexoses. Furthermore, the “paired” fragmentation pattern observed after the Fuc6S residue suggests that the heptose residue could be attached as a branch as indicated by NMR. However, these spectra also contain evidence that the glycan structure is not entirely homogenous. For example, the oxonium ion at m/z 355.1 is likely a hexose–heptose pair. Also, the peak at m/z 856.7 (3+) suggests that an additional hexose can be located between the peptide and the Fuc6S residue.

Properties of the modified Asn residues in Ta0280

We could not obtain an accurate mass for the intact Ta0280 by mass spectrometry, because Ta0280 was not purified to homogeneity and the presence of 1% Triton caused interference. We estimate that the glycan portion represents 48.5 kDa and an observed mass by SDS–PAGE analysis of the mass of Ta0280 considering a calculated mass of 101.5 kDa and an observed mass by SDS–PAGE of 150 kDa. Ta0280 has 93 Asn including 44 that have the Asn–X–Ser/Thr sequon. Since each glycan chain has a mass of 1390 Da, there should be over 30 modified Asn residues. In order to determine if the N-linked glycan is added to specific Asn residues and whether the sites are uniformly glycosylated, we searched the LC-MS and MS/MS data of the in-gel digests (tryptic and chymotryptic) of the main glycoprotein Ta0280 for additional glycopeptides. Seventeen tryptic glycopeptides were identified as well as 12 chymotryptic glycopeptides, representing 14 N-linkage consensus sites. In fact, more chymotryptic glycopeptides were detected but their peptide compositions could not be confirmed from the MS and MS/MS data.

Chymotrypsin cleaves next to aromatic amino acids (Trp, Tyr and Phe) and to a lesser extent at other, neutral residues. Thus, the number of potential chymotryptic glycopeptides that can be produced for Ta0280 is very large, making it difficult to definitively identify them. The glycopeptides identified are presented in Table III and the overall coverage of the Ta0280 amino acid sequence is presented in Figure 7. Though the coverage is not extensive as is often the case for in-gel digests (~50%), it should be noted that all identified glycan residues were associated with peptides that include N-linkage consensus sequons. Furthermore, the majority of the N-linkage sites detected appear to be fully occupied though in some instances, e.g. T426-444 (FSPQKPVMANNSYTLYWK), we find evidence of partial occupancy (data not shown).

### Discussion

The extraction protocol with Triton X-100 in citrate buffer, pH 3, avoided a cell homogenization step to extract the glycoproteins and allowed the quantitative recovery of the glycosylated material in the supernatant after low-speed centrifugation. SDS–PAGE analysis showed that the glycosylated material was composed of both glycoproteins (giving well-defined bands) and material that showed streaking along the lanes by PAS staining. The streaking material could be the lipoglycan that was first reported by Mayberry-Carson et al. (1974) and further characterized by Smith (1980). This lipoglycan was reported to have a “native” apparent mass of over 1000 kDa which was dissociated to polymers with an average

<table>
<thead>
<tr>
<th>Residue</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
<th>H5</th>
<th>H6</th>
<th>H6′</th>
<th>H7</th>
<th>H7′</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) α-galHep</td>
<td>5.576</td>
<td>3.82</td>
<td>3.82</td>
<td>4.20</td>
<td>3.96</td>
<td>3.87</td>
<td>3.91</td>
<td>3.67</td>
<td></td>
</tr>
<tr>
<td>α-Man</td>
<td>97.4</td>
<td>68.8</td>
<td>70.4</td>
<td>68.8</td>
<td>70.6</td>
<td>70.0</td>
<td>64.3</td>
<td>64.3</td>
<td></td>
</tr>
<tr>
<td>(C) α-Man</td>
<td>5.182</td>
<td>4.07</td>
<td>3.88</td>
<td>3.81</td>
<td>3.99</td>
<td>4.04</td>
<td>3.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D) -3,6)-α-Man</td>
<td>103.3</td>
<td>70.9</td>
<td>71.2</td>
<td>69.1</td>
<td>72.5</td>
<td>66.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(E) -3,6)-β-Man</td>
<td>96.9</td>
<td>70.0</td>
<td>81.2</td>
<td>65.8</td>
<td>72.0</td>
<td>66.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(F) -3)-β-Gal</td>
<td>5.054</td>
<td>4.18</td>
<td>3.77</td>
<td>3.94</td>
<td>3.90</td>
<td>4.02</td>
<td>3.77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(G) α-Man</td>
<td>4.993</td>
<td>3.80</td>
<td>3.88</td>
<td>4.24</td>
<td>3.80</td>
<td>3.72</td>
<td>3.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(H) α-Man</td>
<td>80.3</td>
<td>69.1</td>
<td>83.8</td>
<td>69.2</td>
<td>77.1</td>
<td>61.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(I) -4)-β-Glc</td>
<td>4.937</td>
<td>4.01</td>
<td>3.85</td>
<td>3.66</td>
<td>3.69</td>
<td>3.90</td>
<td>3.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(J) -2,3)-β-Fuc6S</td>
<td>100.2</td>
<td>70.8</td>
<td>71.5</td>
<td>67.5</td>
<td>73.7</td>
<td>61.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(K) -1,3)-α-Man</td>
<td>4.918</td>
<td>4.01</td>
<td>3.85</td>
<td>3.66</td>
<td>3.69</td>
<td>3.90</td>
<td>3.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(L) -1,6)-α-Man</td>
<td>104.4</td>
<td>74.5</td>
<td>74.7</td>
<td>77.0</td>
<td>75.4</td>
<td>60.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M) -1,6)-β-Gal</td>
<td>4.695</td>
<td>3.42</td>
<td>3.63</td>
<td>3.80</td>
<td>3.68</td>
<td>3.95</td>
<td>3.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N) -1,6)-β-Man</td>
<td>102.7</td>
<td>70.6</td>
<td>75.5</td>
<td>66.2</td>
<td>71.6</td>
<td>52.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A acetate</td>
<td>0.41</td>
<td>5.57</td>
<td>5.44</td>
<td>5.74</td>
<td>4.45</td>
<td>5.50</td>
<td>4.18/4.14</td>
<td>4.75/4.81</td>
<td></td>
</tr>
<tr>
<td>J acetate</td>
<td>96.7</td>
<td>69.9</td>
<td>71.6</td>
<td>68.6</td>
<td>69.5</td>
<td>67.1/66.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D.E acetate</td>
<td>4.93</td>
<td>3.89</td>
<td>4.37</td>
<td>6.00</td>
<td>4.39</td>
<td>3.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>102.7</td>
<td>74.8</td>
<td>72.6</td>
<td>69.9</td>
<td>72.2</td>
<td>53.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.42</td>
<td>5.41</td>
<td>4.28</td>
<td>5.47</td>
<td>4.28</td>
<td>3.92</td>
<td>4.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>96.8</td>
<td>76.2</td>
<td>70.8</td>
<td>67.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
of 67 kDa in the presence of SDS (Mayberry-Carson et al. 1974). The basic subunit (obtained by the chemical cleavage) was shown to be a linear polymer with the structure \([\text{Man}_\alpha 1,2-\text{Man}_\alpha 1,4-\text{Man}_\alpha 1,3]\text{Glc}_\alpha -\text{diglycerol tetraether}\) (Smith 1980). The chemical and biophysical properties of this lipoglycan would be consistent with the streaking PAS positive material observed on SDS–PAGE.

Mayberry-Carson et al. (1978) reported the detection of glycans on the surface of T. acidophilum using concanavalin A and cytochemical staining. We also observed that the PAS positive material bound to a concanavalin A-sepharose column and was eluted with 100 mM \(\alpha\)-methyl-mannose (data not shown). Both the glycoproteins and the lipoglycan have high mannose content which is consistent with their detection with concanavalin A conjugates.

Using SDS–PAGE and PAS staining, Yang and Haug (1979) reported the presence of two glycoproteins (a major one at 152 kDa and a minor one at 180 kDa) in membranes that were isolated by a sucrose gradient. The 152-kDa glycoprotein that was purified by Yang and Haug (1979) probably
partially purified. It is possible that this glycoprotein was absent in our pool of used a method for glycoprotein purification that differed significantly from the one used by Yang and Haug (1979), since it is smaller than that suggested by Yang and Haug (1979), since it is smaller than the glycan structure that we report here is different from both studies and we used the same T. acidophilum strain [American Type Culture Collection (ATCC) 25905, DSM 1728]. We did not observe a 180-kDa glycoprotein but we the glycan does contain a large proportion of mannose residues. However, in most other respects, the glycan structure that we report here is different from that suggested by Yang and Haug (1979), since it is smaller and contains a FucS residue as well as D-galacto-heptose. Furthermore, the glycan is linked to protein via a Glc-Gal disaccharide and not a chitobiose core as speculated by Yang and Haug (1979). The glycan-peptide linkage via a Glc-Gal appeared to be resistant to PNGase F digestion (data not shown).

It is unclear whether Yang and Haug (1979) analyzed a distinct glycopeptide or whether methodological differences could explain why they proposed a different structure. Our MS/MS data showed some heterogeneity of the glycan structure at specific sites, but these differences imply an alternate linkage of either the heptose or the 6-sulfofucose which are not part of the structure proposed by Yang and Haug (1979). Minor signals observed in the anomeric region of the 1H NMR spectrum (Figure 4) also suggest that some alternate structures could be present. However, these minor signals have intensities that are less than 10% of the anomeric signals of the determined structure (Figure 5) which confirms that our proposed structure is the main one.

**Sulfolobus acidocaldarius**, another thermoacidophilic archaean, was also found to produce glycoproteins extensively modified with an N-linked glycan containing a sulfohexose (Zähringer et al. 2000; Peyfoon et al. 2010). Zähringer et al. (2000) determined that cytochrome b558/566 from *S. acidocaldarius* is glycosylated with a branched hexasaccharide containing 6-deoxy-6-sulfoglucose (6-sulfoquinovose) that is linked to Asn residues via a chitobiose core. Peyfoon et al. (2010) observed that this N-linked glycan is also present at multiple sites on the S-layer glycoprotein of *S. acidocaldarius*. Mapping of the N-linked sequons of the S-layer glycoprotein showed a high glycosylation density, particularly of the C-terminal domain. Heterogeneity of the glycans was

<table>
<thead>
<tr>
<th>Glycopeptide</th>
<th>Amino acid sequence</th>
<th>Observed glycopeptide ions (m/z)*</th>
<th>Glycopeptide mass (Da)</th>
<th>Glycan modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>T&lt;sup&gt;67-82&lt;/sup&gt;</td>
<td>YIGPGSSGNFTIYDAK</td>
<td>973.4 (3+)&lt;sup&gt;*&lt;/sup&gt;</td>
<td>2917.2</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>T&lt;sup&gt;122-139&lt;/sup&gt;</td>
<td>FETADDFVENQTHVINGK</td>
<td>1027.4 (3+)&lt;sup&gt;*&lt;/sup&gt;</td>
<td>3079.2</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>T&lt;sup&gt;308-310&lt;/sup&gt;</td>
<td>LDNITESQLVR</td>
<td>1158.1 (3+)&lt;sup&gt;*&lt;/sup&gt;</td>
<td>3471.3</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>T&lt;sup&gt;311-338&lt;/sup&gt;</td>
<td>YYGGLAQNYTLSTFTNVNYNITSPPK</td>
<td>1104.1 (3+)&lt;sup&gt;*&lt;/sup&gt;</td>
<td>3309.3</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>T&lt;sup&gt;567-596&lt;/sup&gt;</td>
<td>SNNTYLISLSPQSAVSYSFYGGSLTAQK</td>
<td>1285.4 (2+)&lt;sup&gt;*&lt;/sup&gt;</td>
<td>2514.8</td>
<td>Hep, Fuc6S1</td>
</tr>
<tr>
<td>T&lt;sup&gt;826-843&lt;/sup&gt;</td>
<td>AGSYYNATTVWVPPTSGK</td>
<td>1326.0 (4+)&lt;sup&gt;*&lt;/sup&gt;</td>
<td>3529.0</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>C&lt;sup&gt;275-286&lt;/sup&gt;</td>
<td>HLNNGTAFPNYF</td>
<td>1043.1 (3+)&lt;sup&gt;*&lt;/sup&gt;</td>
<td>3126.3</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>C&lt;sup&gt;322-342&lt;/sup&gt;</td>
<td>NITSPPKVTYY</td>
<td>1097.4 (3+)&lt;sup&gt;*&lt;/sup&gt;</td>
<td>3289.2</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>C&lt;sup&gt;452-461&lt;/sup&gt;</td>
<td>VLNNASNAY</td>
<td>1392.9 (2+)</td>
<td>2783.8</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>C&lt;sup&gt;462-471&lt;/sup&gt;</td>
<td>YVYPLNQTVY</td>
<td>1224.0 (2+)</td>
<td>2446.0</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>C&lt;sup&gt;496-503&lt;/sup&gt;</td>
<td>TINGTTEY</td>
<td>1305.0 (2+)</td>
<td>2608.0</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
<tr>
<td>C&lt;sup&gt;505-721&lt;/sup&gt;</td>
<td>KIEYVNGTVAQPSSTY</td>
<td>1350.1 (3+)</td>
<td>3805.2</td>
<td>Hex, Hep, Fuc6S1</td>
</tr>
</tbody>
</table>

*The asterisks (*) indicate the most intense glycopeptide ions.*

It corresponds to the 150-kDa glycoprotein that we identified as Ta0280, since it was found to be the major glycoprotein in two other studies and we used the same *T. acidophilum* strain [American Type Culture Collection (ATCC) 25905, DSM 1728]. It is unclear whether Yang and Haug (1979) analyzed a distinct glycopeptide or whether methodological differences could explain why they proposed a different structure. Our MS/MS data showed some heterogeneity of the glycan structure at specific sites, but these differences imply an alternate linkage of either the heptose or the 6-sulfofucose which are not part of the structure proposed by Yang and Haug (1979). Minor signals observed in the anomeric region of the 1H NMR spectrum (Figure 4) also suggest that some alternate structures could be present. However, these minor signals have intensities that are less than 10% of the anomeric signals of the determined structure (Figure 5) which confirms that our proposed structure is the main one.

*Sulfolobus acidocaldarius*, another thermoacidophilic archaean, was also found to produce glycoproteins extensively modified with an N-linked glycan containing a sulfohexose (Zähringer et al. 2000; Peyfoon et al. 2010). Zähringer et al. (2000) determined that cytochrome b558/566 from *S. acidocaldarius* is glycosylated with a branched hexasaccharide containing 6-deoxy-6-sulfoglucose (6-sulfoquinovose) that is linked to Asn residues via a chitobiose core. Peyfoon et al. (2010) observed that this N-linked glycan is also present at multiple sites on the S-layer glycoprotein of *S. acidocaldarius*. Mapping of the N-linked sequons of the S-layer glycoprotein showed a high glycosylation density, particularly of the C-terminal domain. Heterogeneity of the glycans was
observed with the largest one corresponding to the branched hexasaccharide, while the smaller glycans were presumed to be biosynthetic precursors.

The presence of a sulfohexose in the N-linked glycans of surface glycoproteins from two thermoacidophilic archaeons suggests that it could contribute to cell integrity under these harsh environmental conditions. Because *T. acidophilum* does not display any defined cell wall or S-layer, it is generally considered that the glycan chains of the surface glycoproteins and lipoglycans are directed to the outside of the cells and form a glyocalyx that acts as a protective slime coat (König 1988; Albers and Meyer 2011). The presence of an anionic group on a C-terminal transmembrane domain that allows it to remain associated with the cells. Ta0280, the most abundant glycoprotein, also has a C-terminal transmembrane domain. Even if Ta0280 and Ta1063 are not organized in a clearly defined structure such as the S-layer, their tight association with the surface of the cell and their orientation toward the environment would allow them to play a role in structural integrity.

Abu-Qarn and Eichler (2006) have performed analyses of amino acid sequences surrounding archaeal glycoprotein sequons and found some differences with the sequon-bordering residues present in bacterial and eukaryal glycoproteins. In particular, they reported that the extended sequon characterized by an enhanced presence of Asp or Glu residues at the −2 position in the sequons of *C. jejuni* glycoproteins (Kowarik et al. 2006) was not observed in the confirmed archaeal sequons available when they performed their analysis. We did not observe any Asp or Glu residues at the −2 position of the 14 modified sequons that were characterized in tryptic and chymotryptic digests of Ta0280 (Figure 7) confirming that the extended *Campylobacter* sequon is not essential in archaeal glycoproteins. Abu-Qarn and Eichler (2006) also found that there was a lower probability of aromatic residues at positions −2 and −1 in archaeal sequons (4 and 11%, respectively) when compared with eukaryal sequons. However, we found a moderate frequency of aromatic residues at positions −2 and −1 (21% at both positions) among the 14 modified sequons that we observed in Ta0280.

*Thermoplasma acidophilum* ATCC 25905 (DSM 1728) has a small genome (1.56 Mb) with 1509 predicted open reading frames (Ruepp et al. 2000), including Ta1136, a member of the carbohydrate-active enzymes (CAZy) family GT66 which contains homologs of the β-oligosaccharyltransferase (http://www.cazy.org/GT66_archaea.html). The archaeal β-oligosaccharyltransferase is designated AglB, while the bacterial and eukaryal homologs are designated PglB and Stt3, respectively (Magidovich and Eichler 2009). This enzyme uses lipid-linked donors to transfer oligosaccharides to select Asn residues. Based on further analyses of the glycosyltransferase CAZy families, Magidovich and Eichler (2009) suggested that archaeal systems essentially rely on GT2 and GT4 members to build the oligosaccharides that are transferred by the AglB homologs. The identification of multiple surface glycoproteins in *T. acidophilum* confirms that it has a functional “general” protein glycosylation system that targets proteins that are translocated across the cytoplasmic membrane. The determination of the structure of the N-linked glycan will be useful to decipher the N-glycosylation pathway in this organism and in other archaeans.

**Materials and methods**

**Strain and growth conditions**

We obtained *T. acidophilum* 122-B2 from the ATCC (ATCC 25905, DSM 1728). *Thermoplasma acidophilum* 122-B2 was grown in 22 L of DSM medium 158 (at pH 2.0) but with the initial yeast extract concentration at 0.2% (w/v), in a 28-L fermenter vessel operated at 55°C and with the dissolved oxygen controlled at 10% saturation. A bolus supplementation of yeast extract (0.2%, w/v) was done at 31 h, and 19.3 g
(wet weight) of cell mass was harvested at 50.5 h of growth. The cell mass was kept at −20°C until processing.

**Extraction and purification of the glycoproteins**

*Thermoplasma acidophilum* cells (10 g) were resuspended in 200 mL of 50 mM sodium citrate, pH 3, with protease inhibitor (catalog number P2714, Sigma-Aldrich, St Louis, MO) and stirred at room temperature for 1 h. The cells were centrifuged at 12,000 × g for 10 min. The cell pellet was resuspended in 100 mL of 50 mM sodium citrate, pH 3, with a protease inhibitor, and then mixed with 100 mL of 2% Triton X-100 in the same buffer, to obtain a final concentration of 1% Triton X-100. The resuspended cells were stirred at room temperature for 1 h and then centrifuged at 12,000 × g for 10 min. The 12,000 × g supernatant was centrifuged at 208,000 × g for 1 h and the glycoproteins were recovered in the supernatant. The 208,000 × g supernatant was applied to either a 1-mL Mono S column (for small scale purification) or to a 5-mL HiTrap SP FF cation-exchange column (GE Healthcare, Canada) equilibrated in 20 mM sodium acetate, pH 4, with 0.2% Triton X-100. The column was washed with 20 mL of 20 mM sodium acetate, pH 4, with 0.2% Triton X-100 and 0.2 M NaCl. A gradient of 0.2–1 M NaCl in the same buffer containing Triton X-100 was then applied over a volume of 50 mL. The fractions containing glycoproteins were identified by analysis on SDS–PAGE gel and PAS staining.

**SDS–PAGE, in-gel digest and protein identification**

SDS–PAGE was performed on gels containing 8% acrylamide. Before loading, the samples were incubated for 30 min at 37°C in loading buffer, since we observed that the glycoproteins were better resolved when we avoided heating the samples at higher temperatures. Glycoproteins and other glycoconjugates were stained by the PAS method using the Pierce (Rockford, IL) glycoprotein staining kit (catalog number 24562). For general protein staining, the gels were stained with Coomassie Blue and cut out. N-terminal sequencing was performed on a model 491 Procise protein sequencing system (Applied Biosystems Inc., Foster City, CA) and the bands were stained with Coomassie Blue and cut out. N-terminal sequencing was performed on a model 491 Procise protein sequencing system (Applied Biosystems Inc.) as described by LeGendre and Matsudaira (1988).

**Preparation of glycopeptides using proteinase K digestion**

Protease K (10 mg) was added to a glycoprotein (30 mg) solution in water and kept at 50°C for 2 days. The resulting digest was filtered through a SepPak C18 column (prewashed with methanol and then water) and desalted on a Sephadex G15 column (1.6 × 80 cm) in pyridine (0.4%)–acetic acid (1%) in H2O, monitored by a refractive index detector. All fractions (5 mL each) eluting before the salt peak were dried and analyzed by 1H NMR. Glycopeptide containing fractions were purified by anion-exchange chromatography on a 5-mL HiTrap Q anion-exchange column (GE Healthcare) in a gradient of 0–100% 1 M NaCl over 1 h at a flow rate of 3 mL/min to give oligosaccharides 1a,b.

**Acetylation of the oligosaccharides 1a,b**

Oligosaccharides 1a,b were dissolved in water (0.02 mL) and diluted with pyridine ~1 mL, added slowly to avoid precipitation (oligosaccharides were totally insoluble in pyridine; several attempts were made to keep them in solution). Acetic anhydride (1 mL) was added and the mixture heated for 30 min at 100°C. To ensure complete acetylation, pyridine (3 mL) and acetic anhydride (3 mL) were added and the mixture heated for 1 h at 100°C. Reagents were evaporated with addition of toluene, the product was used for two-dimensional NMR spectra in deuterated acetone without further purification.

**LC-MS analysis**

The in-gel digests were analyzed by nanoLC-MS/MS using a CapLC nanoHPLC system interfaced to a Q-TOF Ultima hybrid quadrupole–time-of-flight mass spectrometer (Waters Corporation, Milford, MA). The samples were injected onto a 300 μm i.d. × 5 mm C18 PepMap100™ trap (LC Packings, San Francisco, CA) and separated on a 75 μm i.d. × 5 cm Picofrit™ column (New Objective, Woburn, MA) using a linear gradient from 5 to 55% acetonitrile+0.2% formic acid (solvent B) over 45 min at a flow rate of ~800 nL/min. Solvent A was 0.2% (vol/vol) formic acid in water. The mass spectrometer was configured to acquire MS/MS spectra on doubly, triply and quadruply charged ions. For protein identification, the MS/MS tandem mass spectra were searched against both the *T. acidophilum* genome and the NCBI nr database using the MASCOT™ search engine (Matrix Science, UK). The search parameters were tryptic or chymotryptic cleavage with up to one missed cleavage site, ≤1.5 Da for precursor ion, ≤1.0 Da for the fragment ions, oxidation (Met) and deamidation (Asn) as variable modifications. Peptide scores >35 were considered as significant. All glycopeptide MS/MS spectra were examined by hand.
NMR analysis

$^1$H and $^{13}$C NMR spectra were recorded by using a Varian Inova 500 or 600 MHz spectrometer in D$_2$O solutions at 40°C with acetone standard (2.225 ppm for $^1$H and 31.5 ppm for $^{13}$C) by using standard pulse sequences COSY, TOCSY (mixing time 120 ms), NOESY (mixing time 200 ms), HSQC and HMBC (long-range transfer delay 100 ms). Selective one-dimensional TOCSY, NOESY and TOCSY–NOESY experiments were used for complete residue assignments and for sequence analysis (Uhrin and Brisson 2000; Brisson et al. 2003). To determine the absolute configuration of the sugars, a conformational analysis was done using the Metropolis Monte Carlo method as described previously (Peters et al. 1993). Calculations were performed for various oligosaccharides with differing absolute configuration up to the complete structure. The molecular model of the octasaccharide was generated using the minimum energy conformer that was compatible with the observed NOEs. Molecular drawings were done using Schakal97 from E. Keeler, University of Freiburg, Germany.

Supplementary data

Supplementary data for this article is available online at http://glycob.oxfordjournals.org/.

Funding

Dr Michel Gilbert acknowledges funding from the National Research Council Canada.

Acknowledgements

We thank Julie Bellefeuille for technical help, Perry Fleming for help with growing T. acidophilum, Luc Tessier for help with mass spectrometry analysis, Dr Malcolm B. Perry for providing a $\alpha$-glycero-$\beta$-galacto-heptose standard, Nam Huan Khieu and Suzon Larocque for help with NMR analysis and Dr Warren Wakarchuk for helpful discussion.

Conflict of interest

None declared.

Abbreviations

Fuc6S, 6-C-sulfo-$\alpha$-fucose or 6-deoxy-6-C-sulfo-$\alpha$-galactose; Hep, $\alpha$-glycero-$\beta$-galacto-heptose; HMBC, heteronuclear multiple bond coherence; HSQC, heteronuclear single quantum coherence; LC-MS, liquid chromatography-mass spectrometry; MS/MS, mass spectrometry/mass spectrometry; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; PAS, periodic acid-Schiff; PVDF, polyvinylidene fluoride; SDS-PAGE, sodium dodecyl sulfate-polycrylamide gel electrophoresis; TOCSY, total correlation spectroscopy.

References


Glycoproteins from Thermoplasma acidophilum