CRYSTAL STRUCTURE OF THE N-TERMINAL DOMAIN OF THE GROUP B STREPTOCOCCUS ALPHA C PROTEIN*

Thierry C. Aupérin§‡†, Gilles R. Bolduc#‡, Miriam J. Baron#, Annie Heroux¶, David J. Filman§, Lawrence C. Madoff#, and James M. Hogle§

§ Harvard Medical School, Dept. of Biological Chemistry and Molecular Pharmacology, 240 Longwood Ave., Boston, MA 02115, # Brigham and Women’s Hospital, Channing Laboratory, Harvard Medical School, Dept. of Medicine, 181 Longwood Ave., Boston, MA 02115, ¶ Biology Department, 463, Brookhaven National Laboratory, Upton, NY 11973-5000

Running Title: Crystal structure of GBS N-terminal domain of Alpha C

‡ Authors contributed equally, || To whom correspondence should be addressed at the § address: Tel: (617) 432-3918; Fax: (617) 432-4360; email: jhogle@hms.harvard.edu, † Current address: Dept. of Biological Sciences, Columbia University, 703 Fairchild Center, M.C. 2452, 1212 Amsterdam Avenue, New York, NY 10027

Group B Streptococcus (GBS) is the leading cause of bacterial pneumonia, sepsis and meningitis among neonates, and an important cause of morbidity among pregnant women and immunocompromised adults. Invasive diseases due to GBS are attributed to the ability of the pathogen to translocate across human epithelial surfaces. The alpha C protein (ACP) has been identified as an invasin that plays a role in internalization and translocation of GBS across epithelial cells. Soluble N-terminal domain of ACP (NtACP) blocks internalization of GBS. We determined the 1.86 Å resolution crystal structure of NtACP comprising residues Ser52 through Leu225 of the full-length ACP. NtACP has two domains, an N-terminal beta sandwich and a C-terminal three-helix bundle. Structural and topological alignments reveal that the beta sandwich shares structural elements with the type III fibronectin fold (FnIII), but includes structural elaborations that make it unique. We have identified a potential integrin-binding motif consisting of Lys-Thr-Asp146, Arg110, and Asp118. A similar arrangement of charged residues has been described in other invasins. ACP shows a heparin-binding activity that requires NtACP. We have validated this prediction using assays of the heparin-binding and cell-adhesion properties of engineered fragments of ACP. This is the first crystal structure of a member of the highly conserved Gram positive surface Alpha-like protein family, and it will enable the internalization mechanism of GBS to be dissected at the atomic level.

Group B Streptococcus (Streptococcus agalactiae, GBS) remains the leading cause of invasive bacterial diseases in neonates, despite its decline in prevalence during the last decade due to intrapartum chemoprophylatic therapy (1,2). It is also an important cause of morbidity in pregnant women and non-pregnant adults with underlying medical conditions (3-5). GBS colonizes the human gastrointestinal and genitourinary tracts, and may cause chorioamnionitis and urinary tract infection in pregnant women and a range of invasive infections in elderly and immunocompromised adults (1,6-8). During labor and delivery, GBS may be transmitted to neonates, causing pneumonia, sepsis or meningitis. Four to six percent of all neonatal GBS infections result in death (6,9).

In vitro, GBS adheres to (10,11), internalizes within (12-14), and translocates across (15) intact human epithelial and endothelial cells. Little is known about the bacterial components that allow this pathogen to adhere to and penetrate cellular membranes. Previous studies suggest that surface proteins are significantly involved in the process (11,16). The surface expressed GBS alpha C protein (ACP) has been shown to act as an invasin (15). ACP is the prototype of a family of surface-expressed proteins containing long tandem repeats (alpha-like proteins, Alp). Members of this family of proteins have a high degree of sequence homology with each other and are thought to share similar function. Indeed, the Alp expressed in group A Streptococcus (GAS), R28, has been associated with cell adhesion (17) and that of
Enterococcus faecalis, Esp, has been associated with virulence (18). Deletion of the ACP gene bca attenuates the virulence of GBS 5- to 7-fold and the internalization of GBS into cervical epithelial cells by 80% (19). Furthermore, the amount of internalized GBS within human cervical epithelial cells in vitro is inhibited by 75% in the presence of the N-terminal domain of ACP (NtACP) (15), suggesting that ACP is a major determinant of virulence. Recently, ACP has been reported to bind glycosaminoglycans (GAG) (20). ACP consists of an N-terminal domain (174 amino acids), a variable number of tandem repeats of 82 amino acids each and a 45-amino acid C-terminal domain containing a LPXTG peptidoglycan-binding motif. NtACP mediates GBS internalization within human epithelial cells (15). NtACP in association with a repeat domain is necessary to bind GAG (20). Moreover, both the alpha C protein and the isolated N-terminal domain are immunogenic and elicit antibodies that protect against GBS infections in experimental animals (21,22).

In order to provide further clues concerning the role of NtACP in adherence, entry, virulence and immunity, we have determined its three-dimensional structure at 1.86 Å resolution by x-ray crystallography. The mostly negative surface of NtACP includes a regular linear arrangement of positive charges that strongly suggested the location of the heparin-binding site, and provided preliminary indications regarding their detailed intermolecular interaction. To confirm these structurally-based predictions, and define the limits of the binding site more precisely, we have conducted in vitro and in vivo binding assays using engineered fragments of ACP. This is the first crystal structure of a member of the highly conserved Gram positive surface Alpha-like protein family, and it will enable the internalization mechanism of GBS to be further dissected at the atomic level.

**MATERIALS AND METHODS**

All amino acid residues in this paper and in the coordinates file are numbered according to the alpha C protein numbering (23). Thus, the first methionine of the recombinant proteins is numbered Met39, as the first residue of the alpha C protein N-terminal domain in the GenBank database is Ser52.

*Cloning and site-directed mutagenesis of NtACP* - Cloning of the DNA encoding for the N-terminal domain of alpha C protein (pDEK14) has been described previously (22). Two leucine residues were substituted with methionine residues (L88M and L94M) using the QuickChange Site-directed Mutagenesis kit (Stratagene) according to the manufacturer’s instructions. Forward primer 5’ - GAT TTA TAT GAT GTA AAA ATG GGT AAA ATA GAT CCA ATG CAA TTA ATT GTT TTA G - 3’, reverse primer 5’ - CTA AAA CAA TTA ATT GCA TTG GAT CTA TTG TAC CCA TTT TTA CAT CAT ATA AAT C - 3’ and pDEK14 template were used. The resulting construct, pGB1, was cloned into Escherichia coli BL21(DE3) for expression of the mutant form of NtACP.

Based on the structure of NtACP, the DNA encoding for D2-R was produced using polymerase chain reaction (PCR) forward primer 5’ - CAC CTT GAG GGA TAA GAT T - 3’ and reverse primer 5’ - TCA TAG TTT ATT TCC TTT ACC - 3’ to amplify a 504 base-pair portion of ACP including approximately one-third of the NtACP (i.e., Leu164 to Leu225, most of domain 2), an adjacent repeat, and 13 amino acid residues of the C-terminal region from plasmid pT7LM39 (24). The forward primer includes a CACC sequence, and the reverse primer includes a TCA stop codon, added as directed for use with the pET200 Directional TOPO Expression Kit (Invitrogen). The PCR product was cloned into the pET200/D-TOPO vector using the manufacturer's instructions, and the resulting construct was transformed into E. coli BL21(DE3).

*Sample preparation* - The native and mutant recombinant NtACP were expressed in E. coli BL21(DE3) (Novagen) with a 6xHis tag fused at their C-terminal end and purified as previously described (15). Selenomethionyl (SeMet) proteins were expressed overnight in E. coli BL21(DE3) at 25°C following a protocol previously described for the use of non-auxotrophic cells (25). SeMet mutant proteins were purified in the presence of 0.05% β-mercaptoethanol. Complete substitution of methionine residues and loss of the N-terminal methionine in native and SeMet-labeled proteins were confirmed by mass spectrometry (MALDI-TOF). Final preparations were dialyzed into 20 mM HEPES pH 7.2, 10 mM DTT.
E. coli BLR (Novagen) was used to express full-length 9-repeat ACP, as well as the 9-repeat region (9RR, without N- or C-terminal regions). Full-length 9-repeat ACP, 9RR and D2-R were expressed and purified as previously described (15, 24).

**Crystallization** - Crystals were grown by hanging drop vapor diffusion method at 21°C using a 2 µl drop. Native and SeMet mutant proteins (18 mg/ml) were mixed separately with an equal volume of the well solution containing 100 mM sodium acetate pH 4.6-5.1 and 10-15% (w/v) PEG 4000 (well solution for SeMet proteins included 10 mM DTT). The resulting elongated crystals (typical dimensions 0.6 x 0.3 x 0.3 mm for the native and 2.0-3.0 x 0.2 x 0.2 mm for the SeMet derivative) grow in space group P6_12_2 with unit cell dimensions of a = b = 56.5 and c = 271.6 Å for the native and a = b = 56.7 and c = 272.2 Å for the SeMet derivative (Table 1). These crystals had one NtACP molecule per asymmetric unit (V_m = 2.88 Å^3/Da) and a solvent content of 57%. In advance of cryo-cooling the crystals, the native crystals were quickly soaked in 100 mM sodium acetate, 10% (w/v) PEG 4000 and 25% glycerol, whereas the SeMet crystals were cryo-protected by bringing the mother liquor to a final concentration of 20% glycerol in small steps (3% increment every 30 min.), and were frozen in liquid nitrogen.

**Data collection** - Data for the native crystals were collected using a 1.01 Å radiation on a nine-element CCD detector at beamline 19ID at the Advanced Photon Source (APS), Argonne National Laboratory. A MAD data set of the SeMet mutant crystal was collected at beamline X26C at the National Synchrotron Light Source, Brookhaven National Laboratory, at three different X-ray energies around the K absorption edge of Se and recorded on an ADSC Quantum 4 CCD detector (Table 1).

**Structure determination and refinement** - The present structure determination depended solely on the anomalous scattering from two selenium atoms (for 180 residues). The crystal structure of SeMet NtACP was determined by the single anomalous dispersion (SAD) method, using its peak wavelength data set (Table 1). The data were integrated and reduced using DENZO and SCALEPACK (26). Atomic positions for two selenium atoms (SeMet^88 and SeMet^94) were located by SOLVE (27) and phases were calculated. A solvent-flattened map (RESOLVE) at 2.60 Å resolution was of high quality (Fig. 1A). Clearly defined solvent boundaries were present and most of the side chain densities were visible. Model building was performed using XtalView (28) and the atomic model was refined using REFMAC5 (29), with experimental phases included as restraints, with intermittent model rebuilding. The resulting model for residues 60-231 with most side chains present gave an R factor and R_free of 26.9% and 30.9%, respectively versus the SeMet data (Table 1). A rigid-body fit of the SeMet model to the native data was performed by molecular replacement (MR) using EPMR (30) at 4 Å resolution. Further refinement of the atomic model and phase extension to 1.86 Å resolution were done with REFMAC5 and additional model rebuilding was performed with XtalView. Solvent molecules (water, DTT, and glycerol) were built based on expected hydrogen bond geometry and electron density. The final R factor and R_free of the NtACP (Fig. 1B) were 19.3% and 23.3%, respectively at 1.86 Å nominal resolution (Table 1). Analysis using PROCHECK (31) shows 97.1% of non-proline and non-glycine residues in the most favored region and 2.85% in the additionally allowed region of the Ramachandran plot. However, one residue (Asn^76) could not be built into the electron density envelope in a single low-energy conformation, suggesting that multiple conformations of the main chain may be present.

**Dot blot assay** - Protein binding to heparin was studied with a modified dot blot technique (20) modeled after an immunoblot assay. Proteins (1 µg each) were applied to a nitrocellulose membrane, followed by blocking for 1 hour with 5% skim milk. The membrane was then incubated with heparin-albumin-biotin (Sigma) at 0.05 mg/ml for 1 hour, followed by alkaline phosphatase-conjugated avidin (Pierce Endogen) for 1 hour prior to washing and developing.

**Flow cytometry** - Fluorescent labeling of proteins was performed using the AlexaFluor 488 Protein Labeling Kit (A-10235; Molecular Probes), to conjugate AlexaFluor 488 dye to protein samples, according to manufacturer’s instructions as previously described (15).

ME180 human cervical epithelial cells (ATCC) were grown to monolayer confluence in 6-well plates with 2 ml of RPMI 1640 ( Gibco), including 10% FCS (Gibco) and 1%
penicillin/streptomycin (Gibco). The day prior to the assay, the medium was replaced with 1 ml of fresh medium and the cells were incubated overnight at 37°C with 5% CO₂. The next day, AlexaFluor-labeled protein was added to the wells to a final concentration of 0.1 µM. The plates were incubated at 37°C with 5% CO₂ for 1.5 hours. The medium was removed from the wells, and the monolayers were washed three times with 1 ml of PBS to remove unbound proteins. 350 µl of trypsin-EDTA (0.25% trypsin, 1 mM EDTA-4 Na, Gibco) was added to the wells, and plates were incubated for 1.5 hours. The medium was removed from the wells, and the monolayers were washed three times with 1 ml of PBS to remove unbound proteins. 350 µl of trypsin-EDTA (0.25% trypsin, 1 mM EDTA-4 Na, Gibco) was added to the wells, and plates were incubated for 10 minutes at 37°C. Cells were detached by repeated pipetting and harvested by centrifugation at 650 rpm (50 x g) for 8 minutes. Cells were washed with 1 ml of PBS and resuspended in 0.1 ml of 2% paraformaldehyde in PBS, before incubation at 4°C overnight. The samples were washed with 1 ml of PBS to remove the fixative, resuspended in 0.4 ml of PBS, filtered through a cell-strainer cap (Falcon), and analyzed by flow cytometry on the MoFLo (Cytomation) machine. The cell population of interest was identified by using the AlexaFluor-labeled BSA sample to define non-specific staining and/or autofluorescence levels. Positive cells were defined to have a fluorescent signal greater than that of 98.5% of the BSA-treated control cell population.

RESULTS

Overall structure - The crystal structure of the NtACP, residues 52-225, has been determined to 1.86 Å resolution by single anomalous dispersion (Table 1 and Fig. 2). The final model has acceptable geometry, with an R factor and R_free of 19.3% and 23.3%, respectively (Table 1). The molecule is elongated with overall dimensions of 82 x 34 x 27 Å. It is composed of two major domains: the membrane-distal domain 1 (D1) is comprised of Ser⁵⁷-Asp¹⁶⁰, and domain 2 (D2) includes Ser¹⁶¹-Leu²²⁵. Domain 1 contains eight beta strands arranged into three beta-sheets: a large flat sheet containing strands F-E-G-I and two smaller sheets, A-D and B-C (Fig. 2). These sheets enclose a hydrophobic core that is highly conserved among the Alp family members (Fig. 3). The F-E-G-I and the A-D sheets are arranged to form a typical beta sandwich, but the basic pattern is elaborated in a distinctive way by the addition of an alpha-helix (H₁, 6 residues) and two 3₁₀-helices (H₂, 4 residues, and H₃, 5 residues). Domain 2 is composed of three antiparallel alpha-helices (H₄, H₅, and H₆) arranged in a ~35 Å long left-handed three-helix bundle, a very common structural motif, that delimits a highly conserved hydrophobic core (Fig. 2 and Fig. 3). The three alpha-helices are composed of 16 amino acids (Ser¹⁶¹ to Asn²⁴⁶), 22 amino acids (Thr¹⁸¹ to Asn²⁰²), and 17 amino acids (Thr²⁰⁵ to Lys²²¹), respectively. An additional, though biologically irrelevant, alpha-helix (H₇, 9 residues), corresponding to the linker from the pET24a vector connecting the protein to the 6xHis tag, immediately follows domain 2.

Structural alignment - Database searches were carried out to establish how domain 1 of NtACP might be related to other proteins of known sequence and structure. Except for the Alps, homology searches at the nucleotide and amino acid levels did not identify any other proteins with meaningful similarities to NtACP. The structure comparison program DALI (32) found no protein that aligned well with domain 1 in its entirety. However, several protein structures align with an rmsd < 3.0 Å with a subset of the alpha carbons comprising 70% of the NtACP domain 1, which includes the F, E, G, and I strands of the large beta-sheet and the A and D strands of the smaller sheet of the beta sandwich. The most significant structural homologues include several proteins known to be involved in cell adhesion, including MADCAM-1 (PDB # 1bqs (33)), type III fibronectin domain of tenascin (PDB # 1qr4 (34)), and module 10 of type III human fibronectin (FnIII10; PDB # 1fnf (35)). The closest structural similarity is seen when the large beta-sheet F-E-G-I of NtACP is aligned with the corresponding beta-sheet C'-C-F-G of FnIII10 (rmsd of 1.1 Å on alpha carbons (Fig. 4A)).

The relationship between the two proteins is made clearer by protein folding topology diagrams of FnIII10 and NtACP domain 1 (Fig. 4B) which were calculated using the TOPS algorithm (36). The diagrams confirm the topological as well as structural correspondence between the large beta-sheet of NtACP and FnIII10 and the correspondences of the A and D strands of NtACP with the A and B strands of the smaller three-stranded sheet of FnIII10. Interestingly, in NtACP a short helix (H₃) occupies the same position in the structure and folding pattern as the third strand of the small sheet of FnIII10 (strand
Crystal structure of GBS N-terminal domain of Alpha C

E) (Fig. 4). NtACP domain 1 is significantly different from other proteins of known structure, due to the presence of a single large and elaborate insertion into the common fold. In the three-dimensional structure of the protein, the inserted elements, \( \beta \)-strands B and C and helices \( H_1 \) and \( H_2 \), are clustered together on one face of the molecule (Fig. 4A). The sequence alignment of the N-terminal domains of Alp family members shows that this large insertion is present and highly conserved in the family members (Fig. 3). The conservation and the pronounced clustering of this large insertion suggest that it may play an essential functional role.

Despite the significant structural differences between NtACP and FnIII10, NtACP contains a large highly exposed loop between the G and I strands of the large sheet that is topologically and structurally related to the integrin-binding loop of FnIII10. In FnIII10, this loop contains a canonical Arg-Gly-Asp (RGD) integrin-binding motif, in which the Arg and Asp side chains are highly exposed. In NtACP, the structurally analogous residues are Lys\(^{144} \) Thr\(^{145} \) Asp\(^{146} \) (KTD). The Lys\(^{144} \) and Asp\(^{146} \) side chains are exposed and are highly conserved in the Alp family. Although this sequence has not been previously described as an integrin-binding sequence, variations on the RGD motifs have been described for other-integrin binding proteins, raising the possibility that integrins may serve as ligands for NtACP-mediated cell adhesion.

Distribution of surface charges - The NtACP surface is dominated by acidic residues, except for two small positively charged clusters (Fig. 5). The distribution of positive charges may be important for understanding the structural basis of the ability of ACP to bind heparin (20). One basic cluster (BR1) is located in domain 1. It includes three basic residues from the vicinity of the E-F loop, Arg\(^{110} \), Lys\(^{114} \), and Lys\(^{106} \), and two amide side chains, Gln\(^{121} \) and Asn\(^{133} \). The positively charged atoms typically lie about 8-9 Å apart and only two of them, Arg\(^{110} \) and Lys\(^{114} \) show charge conservation among Alp family members (Fig. 3 and Fig. 5A). The second, much larger, cluster (BR2) includes basic residues Lys\(^{72} \), Lys\(^{96} \), Arg\(^{165} \), Arg\(^{172} \), Arg\(^{185} \) and the amide side chain of Asn\(^{176} \) in domain 2. These residues are aligned down the cleft between \( \alpha \)-helices H4 and H5 and exposed to the solvent, forming a well-defined positively-charged band of ~35 Å in length (Fig. 5B). Residues included in BR2 are on average 6 Å apart and most of them are conserved (Lys\(^{72} \), Arg\(^{165} \), Lys\(^{196} \), and Arg\(^{185} \)) or conserved by charge (Lys\(^{90} \)) within the Alp family (Fig. 3).

Heparin-binding activity of ACP - Prior work has shown binding of full-length 1-repeat and 9-repeat ACP to ME180 cells in culture and to heparin in a dot blot assay (20). It was reported that the NtACP and the 9-repeat region (9RR), taken separately, bind only minimally in these assays, and that both the NtACP and the repeat-region domains are required for the GAG-binding activity of ACP (20). The NtACP crystal structure shows a cluster of positively-charged residues (BR2) in the three-helix bundle that is predicted to lie near the junction of NtACP with the first of the ACP repeats.

To test whether the BR2 region contributes to the GAG-binding activity of full-length ACP, we expressed and purified D2-R, comprised of the three-helix bundle of the NtACP (starting at Leu\(^{164} \)), one 82-amino acid repeat, and 13 amino acids in the C-terminal region. It thus includes most of the charged residues in BR2, excluding only Lys\(^{72} \) and Lys\(^{90} \), due to their location distant from D2 in the linear sequence of NtACP. D2-R was tested for heparin-binding activity by dot blot analysis as described in (20). It showed a level of activity similar to that of full-length 9-repeat ACP (Fig. 6). In contrast, neither NtACP nor the 9RR region alone bound to heparin appreciably.

D2-R was also evaluated for its interaction with ME180 human cervical epithelial cells in flow cytometry assays performed as described previously (20). We found that D2-R associates with 91.6% of these cells, which is comparable to the binding activity of full-length ACP. Furthermore, this binding was inhibited in a concentration-dependent way in the presence of soluble heparin, as was previously seen for ACP (20). In marked contrast, fewer than 5% of the cells bound either the NtACP or 9RR construct alone (20).

DISCUSSION

Adhesion to human epithelial and endothelial cells is a critical step for GBS colonization and invasion. ACP binds to human epithelial cells and plays a role in the internalization and translocation of GBS across epithelial cells (15). Similarly,
GAS Alp protein R28, identical to GBS Alp3, binds human epithelial cells (17,37). ACP expressed by GBS strain A909 contains a 56 amino acid residues signal sequence at the N-terminus, highly conserved among Alps. It is followed by a 174-residue N-terminal domain (Ser52 to Leu225), nine 82-residue tandem repeats, and a 45-residue C-terminal domain containing a LPXTG peptidoglycan-binding motif. Immunogenic and protective epitopes have been mapped to both the NtACP and repeat regions (22). The number of tandem repeats in ACP expressed by other GBS strains varies from one to sixteen, affecting the antigenicity and presumably the protein structure (38). NtACP competitively inhibits binding of ACP to and internalization of GBS within the human cervical epithelial cell line ME180 (15), suggesting that the N-terminal region of ACP is involved in binding to one or more receptors on the surface of these cells, and that binding facilitates the internalization of GBS. The N-terminal regions of Alps share 60 to 100 percent similarity at the amino acid level (Fig. 3), suggesting that they may function similarly (17,23,37,39).

We have solved the three-dimensional structure of NtACP of GBS strain A909 by x-ray crystallography at high resolution to further understand the role(s) that the N-terminal region of Alps may play during streptococcal infection. NtACP is composed of a membrane-distal beta sandwich at its amino end, and a three-helix bundle at its carboxyl end. In the complete ACP, the carboxyl domain links to the first of the series of repeat regions. Highly conserved residues on the NtACP surface were identified by combining NtACP structural information with a structure-based sequence alignment of the N-terminal regions of the Alp family members. A particularly high degree of sequence conservation was seen along one side of a cleft on top of the molecule, in the G-I loop. The conservation of these residues does not appear to be essential for maintaining the integrity of the three-dimensional fold, which suggests that they may play a functional role instead.

**Potential integrin binding site** - A number of proteins involved in cell adhesion or which serve as cell surface receptors share a common folding pattern. A close structural homology between NtACP beta sandwich (domain 1) and FnIII10 (35) was identified (Fig. 4), wherein each secondary structural element identified as analogous in a topological alignment was also structurally similar. The integrin-binding activity of FnIII10 requires the presence of an Arg1493-Gly1494-Asp1495 (RGD) sequence motif (40,41) on an exposed F-G loop of the structure (35). Although NtACP lacks such an RGD sequence, there is an exposed Lys144-Thr145-Asp146 (KTD) motif at the top end of domain 1. The KTD motif of NtACP is part of the G-I loop that structural and topological alignment procedures have identified as analogous to the loop containing the RGD motif of FnIII10 (35,40). While integrins generally bind RGD sequences, several proteins have been shown to bind a number of related motifs, such as KTS (42,43), MLD (44), or MGD(W) (45) of disintegrins obtustatin, EC3, and EMF-10, respectively. Similarly, COL15, the largest collagenous domain of type XVII collagen, binds to α5β1 and α6β1 integrins through a KGD motif (46). Sequence alignment of the N-terminal domains of the members of the Alp family shows that Lys144 and Asp146 are conserved in all members, and although Thr145 is not conserved, other members display uncharged residues with small side chains (Fig. 3). The conserved residues surrounding the RGD loop of FnIII10 are acidic, as are those surrounding the KTD loop of NtACP.

In addition to the RGD motif of FnIII10, FnIII also requires module 9 for optimal integrin binding affinity (47). Two charged residues in module 9 (FnII9) have been shown to be critical for binding to α5β1 integrin. These are Arg1379, which belongs to a synergy domain with basic residues (35,48,49), and Asp1373 (35,50). Both Arg1379 and Asp1373 are located approximately 30 Å away from the RGD motif of FnIII10 (Fig. 7A). Invasin from Y. pseudotuberculosis (Inv497) is a non-RGD protein that competes with FnIII by binding to α5β1 integrin at the same site as fibronectin (51-53) with about 100 times greater affinity. In reporting the invasin structure, Hamburger et al. (51) proposed that Inv497 (PDB #1cww), though structurally dissimilar from FnIII, shares with FnIII a structurally conserved triangular arrangement of charged residues (Fig. 7A-B). The proposed integrin-binding triangle in Inv497 includes Asp811 and Asp813, which were identified by mutational analysis (54-56), and Arg883, which was suggested by structural considerations (51).

We have found that the structure of NtACP
has a very similar arrangement of exposed aspartate and arginine residues (Fig. 7C), wherein the proposed KTD sequence of NtACP and the known RGD sequence of FnIII10 occupy analogous positions, and NtACP residues Arg\textsuperscript{110} and Asp\textsuperscript{118} are each positioned within 30 Å of the KTD motif, and are structurally analogous to Arg\textsuperscript{1379}, and Asp\textsuperscript{1373} of FnIII9-10. We hypothesize that NtACP and \textit{Y. pseudotuberculosis} invasin may mimic the integrin-binding activity of FnIII in similar ways. Further analysis will be necessary to determine which of the integrins or other molecules of the extracellular matrix (ECM) are bound by NtACP.

**Potential heparin/heparan sulfate binding site** - Several bacterial pathogens express adhesins that bind heparin (57-59). Heparin, an anticoagulant GAG, is composed of repeating disaccharide units of D-glucosamine and uronic acid linked by α(1→4) interglycosidic bonds. Several hydroxyl groups of these monosaccharide residues become sulfated giving rise to a highly negatively charged polymer. These negatively charged sulfates in GAG molecules bind to positively charged groups on proteins (60-62). It has previously been shown that ACP binds to host cell GAG, \textit{in vitro}, in a dose-dependent manner, suggesting the presence of a specific binding site on the surface of the protein (20). High-affinity binding requires the presence both of NtACP and at least one of the ACP repeat domains, suggesting that the complete binding site spans both domains. Knowledge of the detailed molecular structure of NtACP now provides the opportunity to identify which surface residues participate in forming the heparin binding site(s), as a way to begin to understand the molecular basis of infection.

Pure preparations of homogeneously sulfated heparin have well-determined structures forming helical ribbons with about a 5 Å regular spacing of the sulfates along the edge of the ribbon and about 11 Å across the ribbon (63). The binding of heparan sulfate by proteins of known structure can involve a distribution of positively charged side chains on the protein surface that is complementary to the arrangement of the negative sulfate groups along one or both edges of the polymer (64,65). To see if such a regular arrangement of positive charges would be recognizable in NtACP, the distribution of charges on the molecular surface was calculated (Fig. 5). The surface of NtACP is dominated by acidic residues, except for two small positively charged clusters, BR1 and BR2. Residues of BR1 belong exclusively to domain 1 of NtACP, and exhibit an 8-9 Å spacing between residues. BR2 spans both domain 1 and domain 2 of NtACP, and the average spacing between residues is ~6 Å.

Two factors argue that the basic residues of BR2, rather than BR1, are the most likely to be responsible for the NtACP contribution to heparin-binding activity. First, heparin binding requires at least one of the ACP repeats to be present (20). The crystal structure shows that some of the charged residues in BR2 (but none of the residues in BR1) belong to domain 2 and are located quite close to the carboxyl end of NtACP, where the first of the repeats is attached in the full-length ACP. It is easy to imagine a spatially contiguous binding site that includes residues from both domains of NtACP and residues from one or more of the repeat domains as well. A second important factor is that minor changes in the conformations of flexible Lys and Arg side chains could easily reduce the average 6 Å spacing between positive charges that was seen in the crystal structure to the 5 Å spacing required for optimal complementarity. Flexibility in basic side chains of the protein may well favor GAG binding by compensating for heterogeneity in the sequence, conformation, or sulfation pattern of the ligand. The heparin-binding activity of D2-R, which includes the majority of the residues of BR2, supports our hypothesis that the BR2 region of NtACP contributes to the protein's heparin-binding activity. Furthermore, fluorescently labeled D2-R binds to epithelial cells and this binding is inhibited by soluble heparin. These data, in conjunction with the absence of heparin-binding activity in the isolated NtACP or repeat-region constructs alone, suggest that a GAG-binding domain involves a junctional epitope including residues in the BR2 domain as well as residues in the adjacent repeat-region.

The relevance of heparin binding to the process of cell entry by GBS is strongly suggested by the correlation between results in the dot blot and cell-binding assays. Specifically, both samples that showed similarly high levels of binding activity in the dot blot assay also showed high levels in binding epithelial cells. Conversely, both ACP fragments showing negligible levels of binding did so in both assays.

The binding of ACP to cells, and subsequent
steps in cell entry, could involve one or more families of receptor molecules bound simultaneously, alternatively, or in sequence. Possible binding sites for integrin and GAG have been identified, and it is interesting to note that the proposed binding sites for the two ligands lie on opposite faces of the protein. Knowledge of the high-resolution structure of the distal end of ACP provides a framework for mutational studies to identify cellular receptors and their binding sites on the ACP surface. Structural determination of the repeat regions would clearly contribute to further characterize the heparin-binding domain of ACP.

REFERENCES

Crystal structure of GBS N-terminal domain of Alpha C


Crystal structure of GBS N-terminal domain of Alpha C


FOOTNOTES

* The authors would like to thank the staff of APS beamline 19ID (S. Ginell) and the staff of NSLS beamline X26C (S. Myers) for their tremendous help during data collection. We would like to acknowledge the staff of NSLS beamline X12C (R. Sweet and A. Saxena), and the staff of BioCARS (R. Henning) for assistance in developing the crystallization and freezing protocols, and during attempts to find isomorphous derivatives. We thank the members of the Hogle lab (especially L. Guogas, B. Appleton and V. Weiss) as well as members of the Ellenberger lab (especially E. Toth and B. Eichman) at Harvard Medical School for their help and discussion during this work. We would like to thank Meghan Gilmore for assistance in preparing D2-R and Hope Hamrick for assistance with flow cytometry assays. This research was supported in part by PHS grant AI38424 (L. Madoff) and the William Randolph Hearst Fund (G. Bolduc).

a The abbreviations used are: GBS, group B Streptococcus, NtACP, N-terminal domain of alpha C protein (ACP), Alp, alpha-like protein, GAG, glycosaminoglycan, SeMet, selenomethionine(yl), BR, binding region, RGD, Arg-Gly-Asp, KTD, Lys-Thr-Asp, FnIII, type III fibronectin
b The nucleotide sequence for the alpha C protein gene has been deposited in the GenBank database under GenBank Accession Number M97256 (23). The amino acid sequence of this protein can be accessed through NCBI Protein Database under NCBI Accession # AAA26848 (23). The atomic coordinates and structure factor files for the N-terminal domain of Alpha C protein have been deposited with the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/pdb), under PDB # 1ywm (this paper).
FIGURE LEGENDS

Fig. 1. Experimental and final refined electron density maps. The end of the $\beta$-sheet I is shown here. Tyr$_{159}$ is buried between the two domains, with Lys$_{72}$ and Arg$_{165}$ closing the pocket, enclosing two water molecules. (A) The refined SeMet model is shown in stick form on the initial experimental solvent flattened map at 2.60 Å resolution, contoured at 1 $\sigma$. (B) The refined native structure is shown in stick form on the 2Fo-Fc electron density map at 1.86 Å resolution, contoured at 1 $\sigma$. Figures were visualized with XtalView (28) and drawn with RASTER3D (66).

Fig. 2. Stereo ribbon representation of the structure of the N-terminal domain of Streptococcus agalactiae Alpha C protein. $\beta$-sheets are shown in burgundy, $\alpha$-helices in blue, and $3_{10}$-helices in green. Figures were drawn with RIBBONS (67) and rendered with POV-Ray.

Fig. 3. Sequence alignment of the N-terminal domain of the Alp family members. The N-terminal domain sequences of ACP (NCBI Accession # AAA26848) (23), Rib (NCBI Accession # AAC44468) (39), R28 (NCBI Accession # AAD39085) (17), Alp1 (NCBI Accession # AAR08144, Unpublished data), Alp2 and Alp3 (NCBI Accession # AAG01392 and # AAG02097, respectively) (37), and Alp4 (NCBI Accession # CAD32934) (68) are aligned. Highly conserved residues are shaded in black and gray. The secondary structure elements of NtACP are shown as arrows ($\beta$-strands), bars (blue and green, $\alpha$-helices and $3_{10}$-helices, respectively), and lines (connecting loops). The sequences were aligned with ClustalW (69).

Fig. 4. Structural alignment of NtACP domain 1 vs. human type III fibronectin module 10. (A) The NtACP and FnIII10 backbones are superimposed in stereo. (B) TOPS (36) protein folding topology cartoons of FnIII10 (left) and NtACP domain 1 (right). Beta strands are represented as triangles and $\alpha$-helices as circles. The NtACP fold can be understood as a modification and elaboration of the FnIII pattern. The topologically similar secondary structural elements (shown in gray and blue for FnIII and NtACP, respectively) are structurally similar as well. Elements in red occupy analogous positions. Structural elements that are unique to NtACP (green) are found in a single large insertion. Pink arrows indicate the position of the known integrin-binding motif RGD and the putative integrin-binding motif KTD of FnIII10 and NtACP, respectively. Stereo figures were drawn with RIBBONS (67) and rendered with POV-Ray.

Fig. 5. Molecular surface representation of the N-terminal domain of alpha C protein. The views are related by a rotation of 180° about the vertical axis. The structure exhibits a highly acidic surface (red). However, we have identified two positively charged clusters, BR1 and BR2 (blue), as potential glycosaminoglycan-binding sites. Figures were drawn with GRASP (70).

Fig. 6. Localization of the heparin binding activity of ACP by dot blot assay. (a-e) Constructs corresponding to various portions of ACP. (a) Full-length ACP includes NtACP (black), a series of tandem repeats (gray) and a C-terminal domain (white). (b) NtACP alone (c) 9 repeats (9RR). (d) D2-R is responsible for most of the heparin-binding activity of full-length ACP.

Fig. 7. A structurally conserved triangle of exposed charges may be implicated in integrin binding. Ribbon representations of (A) FnIII9-10 (35), (B) Inv497 D4-D5 (51), and (C) domain 1 of NtACP (this paper). Distances reported are measured between alpha carbons. Residues (RG)D$_{1495}$, D$_{373}$, and R$_{1379}$ of FnIII are known to be required for integrin binding (35,40,50). Residues D$_{811}$ and D$_{811}$ of Yersinia Inv497 are known to be involved in integrin binding (54-56,71) and residue R$_{883}$ has been proposed as a structural analogue of R$_{1379}$ of FnIII based on inter-residue distances (51).
### Table 1. Data collection and refinement statistics.

<table>
<thead>
<tr>
<th>Data collection</th>
<th>Native</th>
<th>Mutant SeMet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P6$_1$22</td>
<td>P6$_1$22</td>
</tr>
<tr>
<td>Unit cell parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a$, $b$, $c$, Å</td>
<td>56.46, 56.46, 271.62</td>
<td>56.97, 56.97, 272.23</td>
</tr>
<tr>
<td>$\alpha$, $\beta$, $\gamma$</td>
<td>90, 90, 120</td>
<td>90, 90, 120</td>
</tr>
<tr>
<td>Molecules per asymmetric unit</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Wavelength, Å</td>
<td>1.0100</td>
<td>0.9763</td>
</tr>
<tr>
<td>Resolution range, Å</td>
<td>40-1.86 (1.93-1.86)</td>
<td>30-2.6 (2.69-2.60)</td>
</tr>
<tr>
<td>Compleness, %</td>
<td>95.1 (71.7)*</td>
<td>97.4 (99.2)</td>
</tr>
<tr>
<td>Observations total/unique</td>
<td>267,765 / 19,465 (1,462)</td>
<td>110,146 / 8,581 (828)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>14.0</td>
<td>12.8</td>
</tr>
<tr>
<td>$R_{	ext{sym}}$, %</td>
<td>5.3 (33.9)</td>
<td>6.8 (13.4)</td>
</tr>
<tr>
<td>$&lt;</td>
<td>F_o</td>
<td>/</td>
</tr>
<tr>
<td>Heavy atoms sites (Se)</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

**Refinement**

| Ordered amino acid residues             | 180 (S57-H236) | 172 (P60-L231) |
| Ordered non-hydrogen atoms             | 1,410          | 1,300           |
| - of the protein                        | 528            | 51              |
| - of the solvent                        |                |                 |
| Resolution range, Å                     | 39.5-1.86 (1.96-1.86) | 12-2.6 (2.73-2.60) |
| $R_{	ext{cryst}}$, %                   | 19.3 (21.6)*   | 26.9 (32.7)     |
| $R_{	ext{free}}$, %                    | 23.3 (27.5)    | 30.8 (39.5)     |
| r.m.s.d. bond length, Å                 | 0.015          | 0.010           |
| r.m.s.d. bond angle, °                   | 1.406          | 1.552           |

*Highest resolution shell values are given in parentheses.

† $R_{\text{sym}} = 100 \times \Sigma |I_o - \langle I_o \rangle|/\Sigma I_o$, where $I_o$ is the observed intensity of an individual reflection, and $\langle I_o \rangle$ is the mean intensity of that reflection.

‡ $R_{\text{cryst}} = 100 \times \Sigma |F_o - F_c|/\Sigma |F_o|$, where $F_o$ and $F_c$ are observed and calculated structure factors, respectively.

§ $R_{\text{free}}$, $R_{\text{cryst}}$ calculated for a subset of the reflections (10.1% for the native, 9.8% for the SeMet), which were omitted during the refinement and used to monitor its convergence.
Figure 1
Figure 2
Figure 3
Figure 4
Figure 5
Crystal structure of GBS N-terminal domain of Alpha C

Figure 6

---

(a) 

(b) 

(c) 

(d)
Figure 7

(A) Crystal structure of GBS N-terminal domain of Alpha C

(B) Crystal structure of GBS N-terminal domain of Alpha C

(C) Crystal structure of GBS N-terminal domain of Alpha C
Crystal structure of the N-terminal domain of group B Streptococcus alpha C protein
Thierry C. Auperin, Gilles R. Bolduc, Miriam J. Baron, Annie Heroux, David J. Filman, Lawrence C. Madoff and James M. Hogle

J. Biol. Chem. published online March 6, 2005

Access the most updated version of this article at doi: 10.1074/jbc.M412391200

Alerts:
• When this article is cited
• When a correction for this article is posted

Click here to choose from all of JBC’s e-mail alerts

This article cites 0 references, 0 of which can be accessed free at http://www.jbc.org/content/early/2005/03/06/jbc.M412391200.citation.full.html#ref-list-1