Divergent Modes of Glycan Recognition by a New Family of Carbohydrate-binding Modules*□

Received for publication, December 4, 2007, and in revised form, February 13, 2008. Published, JBC Papers in Press, February 21, 2008, DOI 10.1074/jbc.M709865200

Katie J. Gregg 1, Ron Finn, D. Wade Abbott, and Alisdair B. Boraston 2

From Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia V8W 3P6, Canada

The genomes of myonecrotic Clostridium perfringens isolates contain genes encoding a large and fascinating array of highly modular glycoside hydrolase enzymes. Although the catalytic activities of many of these enzymes are somewhat predictable based on their amino acid sequences, the functions of their abundant ancillary modules are not and remain poorly studied. Here, we present the structural and functional analysis of a new family of ancillary carbohydrate-binding modules (CBMs), CBM51, which was previously annotated in data bases as the novel putative CBM domain. The high resolution crystal structures of two CBM51 members, GH95CBM51 and GH98CBM51, from a putative family 95 \( \alpha \)-fucosidase and from a family 98 blood group A/B antigen-specific endo-\( \beta \)-galactosidase, respectively, showed them to have highly similar \( \beta \)-sandwich folds. However, GH95CBM51 was shown by glycan microarray screening, isothermal titration calorimetry, and x-ray crystallography to bind galactose residues, whereas the same analyses of GH98CBM51 revealed specificity for the blood group A/B antigens through non-conserved interactions. Overall, this work identifies a new family of CBMs with many members having apparent specificity for eukaryotic glycans, in keeping with the glycan-rich environment of C. perfringens would experience in its host. However, a wider bioinformatic analysis of this CBM family also indicated a large number of members in non-pathogenic environmental bacteria, suggesting a role in the recognition of environmental glycans.

Carbohydrates have critical functions in numerous biological events, including, for example, the movement and interactions of cells and proteins in animals, the recycling of plant cell wall carbohydrates, and the interactions between hosts and disease-causing organisms. Central to the role of carbohydrates in biological processes are protein-carbohydrate interactions. Non-catalytic carbohydrate-binding proteins (e.g. lectins, antibodies, and transport proteins) and catalytic carbohydrate-active enzymes are finely tuned to recognize particular carbohydrate structural motifs. The information content of glycans is realized through the specificity of non-catalytic carbohydrate-binding proteins (like lectins and antibodies), whereas carbohydrate-active enzymes change the information content and often unlock the energy contained within these molecules.

Carbohydrate-binding modules (CBMs) are a comparatively new class of non-catalytic carbohydrate-recognizing polypeptide that are generally defined by their presence as ancillary modules in larger, multimodular carbohydrate-active enzymes such as glycoside hydrolases, glycosyltransferases, and polysaccharide lyases (1). In the context of these enzymes, the role of CBMs is to specifically bind the carbohydrate substrate and hold the enzyme in proximity to the substrate, allowing catalysis to proceed more efficiently (2). The number of CBM families, which are defined on the basis of amino acid sequence similarity, has grown to the current number of 50 (www.cazy.org). Although the majority of these families are known to have members that recognize primarily plant cell wall polysaccharides, a growing number of families appear to have specificity for animal glycans and appear within the modular structures of carbohydrate-active bacterial virulence factors (3–6). In these cases, the CBMs appear to have a role in recognizing the information content of cellular glycans, which then allows the enzymatic activity of the virulence factor to be appropriately directed. One of the current challenges in this area, which is complicated by the great diversity of eukaryotic glycans, is to determine the specificity, strength, and molecular determinants of these CBM-glycan interactions. In turn, this provides key information about the cellular targets of the entire carbohydrate-active virulence factor, thus providing greater insight into the host-pathogen interaction.

Clostridium perfringens is a ubiquitous Gram-positive bacterium that is capable of causing an array of diseases such as gastroenteritis and gas gangrene in humans and animals. The genus Clostridium (and the species perfringens in particular) is notable for its prolific production of toxins that contribute to its virulence (7, 8). Among its minor toxins are a battery of carbohydrate-active enzymes of the class glycoside hydrolase, which are thought to aid in major toxin delivery, tissue destruction,

* This work was supported in part by a grant from the Canadian Institutes of Health Research and Michael Smith Foundation for Health Research Scholar. To whom correspondence should be addressed: Dept. of Biochemistry and Microbiology, University of Victoria, P. O. Box 3055 STN CSC, Victoria, BC V8W 3P6, Canada. Tel.: 250-472-4168; Fax: 250-721-8855; E-mail: boraston@uvic.ca.

□ This article was selected as a Paper of the Week.

3 The abbreviations used are: CBMs, carbohydrate-binding modules; NPCBM, novel putative CBM; GaINAC, D-\( N \)-acytylgalactosamine.

1 The atomic coordinates and structure factors (codes 2vmg, 2vmh, 2vmi, 2vnr, 2vng, and 2vno) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

2 Supported by a University of Victoria graduate fellowship.

3 The on-line version of this article (available at http://www.jbc.org) contains supplemental Figs. 1 and 2.

The Journal of Biological Chemistry, Volume 283, number 18, pp. 12604–12613, May 2, 2008 © 2008 by The American Society for Biochemistry and Molecular Biology, Inc.
Structural and Functional Analysis of the CBM51 Family

and nutrient harvesting while in the host. The two myonecrotic strains of *C. perfringens* for which the genome sequences are available are remarkable for their content of highly modular glycoside hydrolases (9, 10). Each strain contains in excess of 50 open reading frames encoding putative glycoside hydrolases, many of which are extracellular and rarely smaller than 1000 amino acids in length and frequently contain, in addition to the catalytic module, three or more definable ancillary modules (3). These ancillary modules include CBMs (3, 4), putative protein-protein interaction domains (11), and modules of unknown function. One such unknown module, which occurs quite frequently in the *C. perfringens* glycoside hydrolases as well as in proteins from a variety of other organisms, is the novel putative CBM (NPCBM) domain. NPCBM domains show no amino acid sequence identity to known carbohydrate-binding proteins but are hypothesized by Rigden (12) to be CBMs based on their context (*i.e.* in carbohydrate-active enzymes) and their proposed β-sandwich fold, which is common to CBMs. To test the hypothesized carbohydrate-binding function of these modules, we performed structural and functional studies on the NPCBMs from two different multimodular clostridial enzymes, one a hypothesized family 95 α-L-fucosidase (CPF_2129 from *C. perfringens* strain ATCC 13124) and the other a confirmed family 98 blood group A/B antigen-specific endo-β-D-galactosidase (EabC from *C. perfringens* strain ATCC 10543) (Fig. 1) (13). The two targeted modules share only ~30% amino acid sequence identity. The results confirm the identification of NPCBMs as a new CBM family, now classified as CBM51, and define a new module that can mediate the carbohydrate-binding interactions of the enzymes with host tissues. Furthermore, an unexpected degree of diversity in specificity and binding site architecture is evident in this new CBM family, providing some distinction from other CBM families.

EXPERIMENTAL PROCEDURES

Cloning, Gene Expression, and Protein Production and Purification—The gene fragments encoding GH95CBM51 and GH98CBM51 were PCR-amplified from the genomic DNA of *C. perfringens* strains ATCC 13124 and ATCC 10543, respectively, using the following sets of primers: GH95CBMF (5'-CAT ATG GCT AGC GAA AAG GTT GCA GTT G-3') and GH95CBMR (5'-GAA TTC TCT CTC GAG TTA TGT TAA CTT AGC G-3') for GH95CBM51 and GH98CBMF (5'-CAT ATG GCT AGC GAA GTT TAT GCT TGT GAA GAC G-3') and GH98CBMR (5'-GAA TTC TCT CTC GAG TTA ATT CAC AAA ATC ACC CTT AGC TGT C-3') for GH98CBM51. The amplified products were digested with NheI and XhoI restriction endonucleases and ligated to like digested pET-28a using standard cloning procedures. The resulting plasmids, pGH95CBM and pGH98CBM, encode the desired CBM fused to an N-terminal His6 tag by a thrombin protease cleavage site. Polypeptides were produced in 4-liter cultures of *Escherichia coli* strain BL21(DE3) containing pGH95CBM and pGH98CBM using the methods described previously (3). GH95CBM51 and GH98CBM51 were purified by immobilized metal affinity chromatography from cell-free extracts following previously described procedures (3). Purified polypeptides were concentrated and exchanged into 20 mM Tris-HCl (pH 8.0) in a stirred ultrafiltration unit (Amicon, Beverly, MA) using a 5-kDa cutoff membrane (Filtron Corp., Northborough, MA). Purity as assessed by SDS-PAGE was >95%.

Selenomethionine-labeled GH95CBM51 was produced using the *E. coli* B834(DE3) methionine auxotroph. *E. coli* colonies taken from an LB-agar plate were used to inoculate 3 liters of SelenoMet Medium Base (Molecular Dimensions Ltd.) supplemented with SelenoMet Nutrient Mix (Molecular Dimensions Ltd.) and L-selenomethionine (40 mg/liter). These cultures were grown, induced, and harvested, and the polypeptide was purified exactly as described for the unlabeled protein. Determination of Protein Concentration—The concentration of purified protein was determined by UV absorbance (280 nm) using calculated molar extinction coefficients of 17,420 and 24,410 M⁻¹ cm⁻¹ for GH95CBM51 and GH98CBM51, respectively.

Binding Studies—Proteins for glycan array screening were labeled with fluorescein isothiocyanate (Invitrogen) according to the manufacturer’s directions. Labeled protein was desalted and separated from free fluorescein isothiocyanate by gel filtration chromatography using Sephadex G-25 (GE Healthcare). Fluorescein isothiocyanate-labeled GH95CBM51 and GH98CBM51 were used to probe the printed glycan arrays following the standard procedure of Core H of the Consortium for Functional Glycomics (www.functionalglycomics.org/).

Isothermal titration calorimetry was performed as described previously using a VP-ITC system (MicroCal, Northampton, MA) (3). Protein samples were dialyzed extensively against 50 mM Tris-HCl (pH 7.5) and 1 mM CaCl₂ and then concentrated in a stirred ultrafiltration cell as described above. Sugar solutions were prepared by mass in buffer saved from the ultrafiltration step. Both protein and sugar solutions were filtered and degassed immediately prior to use. Protein concentrations were determined by UV absorbance as described above. Although the concentrations of acceptor (i.e. GH95CBM51 or GH98CBM51) were quite high (between 200 and 500 μM), the low affinities of the interactions resulted in C values <2 (14). Thus, based on the 1:1 binding observed in the crystal structures, the data were fit with a single binding site model using stoichiometries fixed at 1. The association constants (Kₐ) were determined by this fitting process.

Crystallization, Data Collection, and Structure Solution—GH95CBM51 and GH98CBM51, previously exchanged into 20 mM Tris (pH 8.0), were treated overnight at room temperature with thrombin to remove the His₆ tag. The polypeptides were separated from the cleaved His₆ tag by size exclusion chromatography using a Sephacryl S-200 column (GE Healthcare).
Pure fractions were concentrated in a 10-ml stirred ultrafiltration device using a 5-kDa cutoff membrane. All crystallizations were performed by the hanging drop vapor diffusion method at 18 °C. Diffraction data for all crystals except the selenomethionine-substituted GH95CBM51 crystals were collected on our home source: a Rigaku R-AXIS IV detector coupled to an area detector. Crystals of native and selenomethionine-labeled GH95CBM51 (both at 15 mg/ml) were grown in 0.05 M calcium chloride, 0.1 M tris(hydroxymethyl) aminomethane (pH 7.5), and 20% polyethylene glycol 3350. An optimized selenium single anomalous dispersion diffraction data set for selenomethionine-labeled GH95CBM51 was collected on beamline X6A at the National Synchrotron Light Source (Brookhaven National Laboratories). SHELXC/D was selected on beamline X6A at the National Synchrotron Light Source (Brookhaven National Laboratories). SHELXC/D was used to determine the substructure of three selenium atoms, which were located with the molecular replacement program MOLREP. Building and refinement were carried as described above.

Water molecules were added to all models using the REFMAC implementation of ARP/wARP (18) and inspected visually prior to deposition. In all data sets, 5% of the observations were flagged as “free” (22) and used to monitor refinement procedures. All final model statistics are given in Table 1. Structure images were prepared with PyMOL.

Bioinformatic Analysis—Putative CBM51 domains were detected by position-specific iterative BLAST searches (23) using both GH95CBM51 and GH98CBM51 as queries. Polypeptides containing similar amino acid sequences were dissected for modularity by InterProScan (24) and classified using both GH95CBM51 and GH98CBM51 as queries.

Crystals of native and selenomethionine-labeled GH98CBM51 (both at 15 mg/ml) were grown in 0.2 M sodium carbonate, 0.1 M HEPES (pH 7.5), and 20% polyethylene glycol 3350. An optimized selenium single anomalous dispersion diffraction data set for selenomethionine-labeled GH98CBM51 was collected on our home source: a Rigaku R-AXIS IV detector coupled to an area detector. Crystals of GH98CBM51 at 15 mg/ml were grown in 0.2 M sodium carbonate, 0.1 M HEPES (pH 7.5), and 20% polyethylene glycol 3350. The new asymmetric unit of the soaked crystals contained two molecules of GH98CBM51, each bound to carbohydrate, which were located with the molecular replacement program MOLREP. Building and refinement were carried as described above.

Water molecules were added to all models using the REFMAC implementation of ARP/wARP (18) and inspected visually prior to deposition. In all data sets, 5% of the observations were flagged as “free” (22) and used to monitor refinement procedures. All final model statistics are given in Table 1. Structure images were prepared with PyMOL.

Bioinformatic Analysis—Putative CBM51 domains were detected by position-specific iterative BLAST searches (23) using both GH95CBM51 and GH98CBM51 as queries. Polypeptides containing similar amino acid sequences were dissected for modularity by InterProScan (24) and classified when possible based on the predicted enzymatic specificity for an appended catalytic module. In some cases, boundary positions were fine-tuned by direct sequence alignments with ClustalW (25). For enzymes that contained tandem copies of CBM51, each module was treated independently. Sequence entries were cut off at >20% identity over a mini-

---

**TABLE 1**

X-ray data collection and refinement statistics

| Data collection | GH98CBM51 | GH95CBM51 |
|-----------------|-----------|-----------|
| Space group     | P2₁,2₁,2  | P2₁,2₁,2  |
| Asymmetric unit | 2P₂₁,2₁,2 | 2P₂₁,2₁,2 |
| **Resolution (Å)** | 20.00-1.55 | 20.00-1.50 |
| **No. reflections** | 25,390 | 25,390 |
| **Rmerge (%)** | 18.188 | 18.188 |
| **Completeness (%)** | 97.0 (99.4) | 97.0 (99.4) |
| **Redundancy** | 13.3 (2.7) | 13.3 (2.7) |
| **Resolution (Å)** | 20.00-1.45 | 20.00-1.40 |
| **No. reflections** | 57,807 | 57,807 |
| **Rmerge (%)** | 18.22/26 | 18.22/26 |
| **Completeness (%)** | 89.6 (76.3) | 89.6 (76.3) |
| **Redundancy** | 18.6 (23.5) | 18.6 (23.5) |
| **Resolution (Å)** | 20.00-1.40 | 20.00-1.40 |
| **No. reflections** | 55,064 | 55,064 |
| **Rmerge (%)** | 15.0/17.8 | 15.0/17.8 |
| **Completeness (%)** | 95.6 (97.7) | 95.6 (97.7) |
| **Redundancy** | 37.47, 47.95, 38.70 | 37.47, 47.95, 38.70 |
| **Resolution (Å)** | 20.00-1.35 | 20.00-1.35 |
| **No. reflections** | 18,188 | 18,188 |
| **Rmerge (%)** | 1109 | 1109 |
| **Completeness (%)** | 11.8 (15.0) | 11.8 (15.0) |
| **Redundancy** | 132 | 132 |
| **Resolution (Å)** | 20.00-1.30 | 20.00-1.30 |
| **No. reflections** | 1116 | 1116 |
| **Rmerge (%)** | 15.0 (16.0) | 15.0 (16.0) |
| **Completeness (%)** | 11.4 (15.4) | 11.4 (15.4) |
| **Redundancy** | 146 | 146 |
| **Resolution (Å)** | 20.00-1.25 | 20.00-1.25 |
| **No. reflections** | 1100 | 1100 |
| **Rmerge (%)** | 23.8 | 23.8 |
| **Completeness (%)** | 19.7 (22.7) | 19.7 (22.7) |
| **Redundancy** | 30.6 | 30.6 |
| **Resolution (Å)** | 20.00-1.20 | 20.00-1.20 |
| **No. reflections** | 23.8 | 23.8 |
| **Rmerge (%)** | 0.020 | 0.020 |
| **Completeness (%)** | 1.695 | 1.695 |
| **Redundancy** | 0.022 | 0.022 |
| **Resolution (Å)** | 1.881 | 1.881 |
| **No. reflections** | 0.022 | 0.022 |
| **Rmerge (%)** | 1.821 | 1.821 |
| **Completeness (%)** | 1.602 | 1.602 |
| **Redundancy** | 0.019 | 0.019 |
| **Resolution (Å)** | 1.603 | 1.603 |
| **No. reflections** | 0.019 | 0.019 |
| **Rmerge (%)** | 0.937 | 0.937 |
| **Completeness (%)** | 0.92 (2.7) | 0.92 (2.7) |

---

**TABLE 2**

Structural and Functional Analysis of the CBM51 Family

SeMet, selenomethionine; r.m.s., root mean square.

| Structural and Functional Analysis of the CBM51 Family | GH98CBM51 | GH95CBM51 |
|-------------------------------------------------------|-----------|-----------|
| **Enzyme**                                             |           |           |
| GH98CBM51                                             |           |           |
| GH95CBM51                                             |           |           |
| **SeMet, selenomethionine; r.m.s., root mean square.** |           |           |

---

**TABLE 3**

Bioinformatic Analysis—Putative CBM51 domains were detected by position-specific iterative BLAST searches (23) using both GH95CBM51 and GH98CBM51 as queries.
mum of 100 amino acids. Phylogenetic tree construction was performed with PhyloDraw Version 0.8 (26).

RESULTS AND DISCUSSION

A New Family of Carbohydrate-binding Modules—To facilitate the characterization of NPCBMs in isolation, the gene fragments encoding the modules were cloned and over-expressed in E. coli, and the polypeptides were purified by immobilized metal affinity chromatography. The purified polypeptides, here called GH95CBM51 (from C. perfringens strain ATCC 13124 GH95) and GH98CBM51 (from C. perfringens strain ATCC 10543 GH98), were screened for glycan binding by Core H of the Consortium for Functional Glycomics using glycan microarrays. Both proteins displayed the ability to recognize glycans (Fig. 2). GH98CBM51 was clearly the more specific module, showing binding only to glycans bearing the blood group A or B antigen trisaccharide determinants (Fig. 2B). This specificity is in keeping with the blood group antigen specificity of GH98 (EabC) (13). In contrast, GH95CBM51 appeared to interact with a large number of the glycans presented in this array. The common determinant of the best hits was terminal galacto-configured sugars, either D-galactose or D-N-acetylgalactosamine (GalNAc) (Fig. 2A). The exceptions contained galactose following terminal α-2,6-linked sialic acid or α-1,2-linked fucose. GH95 is a predicted α-fucosidase; thus, the binding specificity of its

FIGURE 2. Glycan microarray binding data for GH95CBM51 (A) and GH98CBM51 (B). Glycans giving substantial signals are indicated with their structures. Yellow symbols indicate galacto-configured monosaccharides; blue symbols indicate gluco-configured monosaccharides; and green symbols indicate manno-configured monosaccharides. Circles represent hexose monosaccharides, and squares indicate their 2-acetamido derivatives. Purple diamonds indicate sialic acid, and red triangles represent l-fucose.
CBM, GH95CBM51, is somewhat at odds with the predicted catalytic specificity. The biological significance of this is currently unknown.

The carbohydrate-binding activities of GH95CBM51 and GH98CBM51 were confirmed and quantified by isothermal titration calorimetry. GH95CBM51 bound Gal\(\beta_1-3\)GalNAc and methyl-\(\beta\)-Gal with association constants of \((1.48 \pm 0.05) \times 10^3\) and \((0.48 \pm 0.04) \times 10^3\) M\(^{-1}\), respectively. GH98CBM51 bound the A antigen tetrasaccharide (Gal\(\beta_1-3\)(Fuc\(\beta_1-2\)Gal\(\beta_1-4\)GlcNAc\(\beta\)-CH\(_2\)-CH\(_2\)-N\(_3\)) and B antigen tetrasaccharide (Gal\(\beta_1-3\)(Fuc\(\beta_1-2\)Gal\(\beta_1-4\)GlcNAc\(\beta\)-CH\(_2\)-CH\(_2\)-N\(_3\)) with association constants of \((3.85 \pm 0.25) \times 10^3\) and \((1.30 \pm 0.01) \times 10^3\) M\(^{-1}\), respectively. In general, these affinities are on the lower end of the affinity range for protein-carbohydrate interactions. The higher affinity of GH95CBM51 for the disaccharide ligand Gal\(\beta_1-3\)GalNAc versus the methylated galactose monosaccharide suggests additional interactions with the second sugar residue. GH98CBM51 showed a slight (2–3-fold) preference for the blood group A antigen tetrasaccharide in comparison with the blood group B antigen tetrasaccharide. This is consistent with the smaller observed signal for the blood group B antigen-containing glycans in the microarray binding analysis (Fig. 2B). Overall, these results demonstrate that these two examples of NPCBM domains are indeed carbohydrate-binding modules, and this family of modules has now been reclassified as CBM51.

CBM51 Displays a Common Carbohydrate-binding Fold—We approached the question of what are the molecular determinants of specificity in this family 51 CBMs (NPCBMs) through x-ray crystallography. The x-ray crystal structures of GH95CBM51 and GH98CBM51 were solved by optimized selenium single anomalous dispersion and molecular replacement, respectively. Both modules adopt a \(\beta\)-sandwich fold comprising a five-stranded antiparallel \(\beta\)-sheet opposing a four-stranded antiparallel \(\beta\)-sheet (Fig. 3, A and B). The \(\beta\)-sandwich fold is common to numerous families of carbohydrate-binding proteins, including a number of CBM and lectin families (1). The atomic models of GH95CBM51 and GH98CBM51 are obviously similar and overlap with a root mean square deviation of 1.2 Å over 140 matched residues (measured by the secondary structure matching algorithm (27) as implemented in COOT). The primary difference in the structures of these two polypeptides is a small insertion in GH98CBM1, which results in the addition of a small \(\alpha\)-helix near the binding site (Fig. 3B). Relative to GH95CBM51, this addition in GH98CBM51 appears to play a role in recontouring the binding site to accommodate the larger blood group A/B carbohydrates.

Both polypeptides bind a metal atom (Fig. 3, A and B) that was judged to be Ca\(^{2+}\) based on the strictly oxygen-mediated coordination, coordination bond lengths that range from 2.25 to 2.40 Å, a B-factor when modeled as Ca\(^{2+}\) that is consistent with the neighboring atoms, and a small peak of anomalous signal that overlaps with the position of this atom (data not shown). The position of this atom is conserved between GH95CBM51 and GH98CBM51 (within 0.6 Å in a structural overlap) but is not conserved in other CBM families. In the case of these family 51 CBMs, the metal-binding site is at the interface of the two \(\beta\)-sheets (the edge of the sandwich) and in the

---

**FIGURE 3. Fold and calcium binding of CBM51.** Shown are color-ramped schematic representations of GH95CBM51 with methyl-\(\beta\)-galactose (A) and GH98CBM51 with the B antigen trisaccharide (B), both with bound calcium atoms shown as pink spheres and ligands in stick representation. C shows a schematic representation of CBM6-1 from Clostridium stercorarium (Protein Data Bank code 1uy4) (34), with its bound calcium atom shown as a pink sphere and its bound xylotetraose ligand shown in stick representation. This structure is a representative \(\beta\)-sandwich CBM showing the common calcium-binding site.
loops that join the termini of the β-strands, near the carbohydrate-binding site (Fig. 3, A and B). In contrast, the majority of other CBM families that bind metal atoms, which notably are all also β-sandwich proteins, do so at a site that is also at the edge of the sandwich but on a side such that the binding site is on the edge of the strands rather than at the ends (Fig. 3C).

GH95CBM51 Has a Simple Galactose-binding Site—A ligand-bound form of GH95CBM51 was obtained by co-crystallization of the protein with excess methyl-β-D-galactose. The electron density for this sugar was very clear, allowing unambiguous modeling of the ligand (Fig. 4A).

The binding site of GH95CBM51 is quite shallow (Fig. 4A) and provides surprisingly few direct interactions with the sugar (Fig. 4B). The ε-nitrogens of His955 and His1041 make hydrogen bonds with the sugar O-4 and O-3, respectively (Fig. 4B). The galactose O-4 also makes a hydrogen bond with the backbone nitrogen of Ser1039 (Fig. 4B). The phenol group of Tyr922 makes a classical protein-carbohydrate hydrophobic interaction with the apolar plane formed by C-3, C-4, C-5, and C-6 on the B-face of the galactose. This mode of interaction, which requires an equatorial O-2, an equatorial O-3, and an axial O-4, is consistent with the observed specificity of GH95CBM51 for terminal galacto-configured sugars. However, the galactose O-2 and O-6 are solvent-exposed, suggesting that modifications may be tolerated at these positions. Indeed, the array screening results indicated that this polypeptide can interact with terminal GalNac (i.e. can tolerate an acetamido modification at C-2) and suggested an ability to accommodate a fucosyl residue α-1,2-linked to galactose. Likewise, the exposure of the O-6 group may also explain the apparent binding of GH95CBM51 to a glycan terminating in sialic acid α-2,6-linked to galactose.

The galactose-binding site of GH95CBM51 shows a reasonable degree of architectural similarity to the binding sites of the galactose-specific family 32 CBMs, the best characterized of which is CpbCM32, the CBM32 from C. perfringens GH84C (NagJ) (3). Although GH95CBM51 and CpbCM32 show no significant amino acid sequence similarity, they share the same fold (root mean square deviation of ~3.3 Å) and binding site location (Fig. 4C). Many of the protein-carbohydrate interactions are also conserved. Tyr922 of GH95CBM51 is structurally conserved with Trp661 of CpbCM32, which plays a role in binding similar to Tyr922. His955 is conserved with His658, which, like His955, hydrogen bonds with O-4 of the galactose. In CpbCM32, the N79 of Arg690 provides the same hydrogen bond to O-3 of the galactose as the Nε of His1041 in GH95CBM51. Thus, the majority of the GH95CBM51-galactose interactions are structurally well conserved with the family 32 CBMs, perhaps reflecting a simple but prototypical mode of galactose recognition by CBMs.

Structural Basis of Blood Group Antigen Recognition—Ligand-bound forms of GH98CBM51 with blood group A and B carbohydrates were obtained by crystal soaking experiments. In both cases, the electron density for the carbohydrate ligands was quite clear, facilitating accurate modeling of these sugars (Fig. 4, D and E). Only in the case of the 2-acetamido group of the GalNAc in the blood group A antigen trisaccharide was there any ambiguity, as there was no clear electron density for C-8 of this residue. This likely reflected a certain degree of disorder in this portion of the molecule, as this observation was made for the A antigen trisaccharide ligands in the binding sites of both molecules of GH98CBM51 in the asymmetric unit.

The ligand specificity of GH98CBM51 is substantially more restricted in comparison with GH95CBM51, which is revealed by the architecture of its binding site. It is comparatively deep (Fig. 4, D and E), and specific interactions are made between the protein and all three sugars of the blood group carbohydrates (Fig. 4F). The constellation of interactions between GH98CBM51 and these sugars is identical for both blood groups A and B. The plane formed by C-3, C-4, C-5, and C-6 on the B-face of the central galactose residue makes a hydrophobic interaction with Trp192, whereas O-4 of this sugar hydrogen bonds with O" of Tyr135. Asp56 hydrogen bonds with O-2 and O-3 of the fucose residue, Lys61 with O-4 of this residue, and Thr58 with O-3 of the fucose residue. Specificity for the A/B blood group is provided by hydrogen bonds between Asp195 and the galactose or GalNAc O-4 and between Thr70 and O-3. No direct interactions are evident involving the 2-acetamido group of the blood group A determinant, GalNAc, or involving O-2 of the blood group B determinant, galactose, explaining the general ability of this polypeptide to bind both sugars. However, GH98CBM51 did display a 2–3-fold preference for the blood group A antigen, for which there is no obvious structural explanation (i.e. no differences in direct interactions are evident). However, it appears unlikely that this small preference is biologically relevant.

GH98CBM51 is the only CBM known to be specific for the blood group A/B antigens. Although we did recently determine the structure of a family 47 CBM from Streptococcus pneumoniae in complex with the gluco analog of the blood group A antigen tetrasaccharide, this sugar was not the preferred ligand for this protein (5). In general, structural information regarding the interaction of other carbohydrate-binding proteins with their natural blood group A/B antigens is quite scarce. Only the structures of two lectins and a viral capsid protein have been determined in complex with A and/or B antigen-reactive tri- or tetrasaccharides (28–30). Interestingly, these structures reveal that there is little conservation in the modes of their carbohydrate recognition. The norovirus capsid protein maintains specific interactions between the protein and the terminal Gal/GalNAc that defines the antigen and the fucose residue (29). It makes no interactions with the central galactose residue and does not utilize the classical aromatic amino acid side-chain-sugar ring hydrophobic interaction. Winged bean basic agglutinin I maintains the majority of its interactions, including an aromatic side chain hydrophobic interaction, with the terminal Gal/GalNAc of the antigen but also has some polar interactions with the central galactose residue (30). The fucose residue apparently does not participate in the interaction. Similarly, the fungal galectin CGL2 makes specific interactions with the blood group-determining Gal/GalNAc and the central galactose residue, with which a trypophan side chain makes an apolar interaction (28). The fucose residue is not involved blood group recognition by CGL2. GH98CBM51 is unique, as it makes specific interactions with all of the residues in the blood group A/B antigen trisaccharide (the blood group-determining
FIGURE 4. Carbohydrate-binding sites of GH95CBM51 and GH98CBM51. A, solvent-accessible surface representation of the GH95CBM51 binding site containing methyl-β-D-galactose. The sugar is shown in green stick representation with maximum likelihood (20)/\sigma_A (35) weighted 2F_o − F_c and F_o − F_c electron density maps contoured at 0.13 electrons/Å^3 (1σ) and 0.08 electrons/Å^3 (2.5σ), respectively. B, binding site showing the bound sugar in green stick representation and side chains involved in binding the sugar in blue stick representation. Potential hydrogen bonds are shown as purple dashed lines. C, overlay of the GH95CBM51 binding site (yellow) and the binding site of CpCBM32 (magenta) from C. perfringens GH84C (NagJ). Structurally conserved residues involved in binding are shown in stick representation. The numbering indicates the residues in CpCBM32. The residues shown for GH95CBM51 are the same as those indicated in B. D, solvent-accessible surface representation of the A antigen trisaccharide in the GH98CBM51 binding site. The sugar is shown in green stick representation with maximum likelihood (20)/\sigma_A (35) weighted 2F_o − F_c and F_o − F_c electron density maps contoured at 0.16 electrons/Å^3 (0.9σ) and 0.09 electrons/Å^3 (2.5σ), respectively. E, solvent-accessible surface representation of the B antigen trisaccharide in the GH98CBM51 binding site. Electron density maps are maximum likelihood/\sigma_A weighted 2F_o − F_c and F_o − F_c maps contoured at 0.16 electrons/Å^3 (1.0σ) and 0.08 electrons/Å^3 (2.5σ), respectively. F, binding site showing the bound A antigen trisaccharide in green stick representation and side chains involved in binding the sugar in blue stick representation. Potential hydrogen bonds are shown as purple dashed lines. The pattern of interactions is identical for the B antigen trisaccharide and is not shown.
Gal/GalNAc, galactose, and fucose) (Fig. 4, D–F). The glycan microarray screening did not reveal binding to any fragments of this ligand, indicating that this triumvirate of residues is required for binding. Furthermore, a model of GH98CBM51 with the blood group A antigen tetrasaccharide suggests that the GlcNAc residue of this sugar may even make additional interactions with this polypeptide (data not shown). Although the affinity of GH98CBM51 for the blood group A/B antigens is low, it appears that GH98CBM51 may be one of the most selective carbohydrate-binding proteins for the complete fucosylated forms of these sugars.

**Diversity in the CBM51 Family**—A structural overlap of GH95CBM51 and GH98CBM51 reveals how the binding sites of these two related CBMs are adapted to recognize their different respective ligands. The GalNAc of the blood group A antigen in the GH98CBM51 binding site approximately overlaps with the galactose in the GH95CBM51 binding site (Fig. 5A). Hydrogen bonding to O-3 and O-4 of the sugar is maintained by residues that are roughly structurally conserved, although the nature of the side chains differs substantially: two histidines in GH95CBM51 versus one aspartic acid and one threonine in GH98CBM51. The tyrosine in GH95CBM1 is structurally replaced by an aspartic acid in GH98CBM1, where instead of making hydrophobic interactions with the GalNAc, it is involved in polar interactions with the fucosyl moiety of the blood group antigen. The remaining four binding residues in GH98CBM1 are unique to this module and appear distinctively positioned to accommodate the blood group antigens.

The observed lack of conserved functional residues between these CBMs prompted us to undertake a more comprehensive bioinformatic analysis of the CBM51 family. The CBM51 family currently comprises ~60 entries from ~46 different proteins from ~26 different organisms (different strains of a given species are included in this count). These appear to cluster into six different subfamilies (a–f) (Fig. 5B).
The ammonia identity with the catalytic module of common specificity among members of a given CBM family is reasonably conserved among these subfamilies. Thus, the degree of divergence between subfamilies CBM51a and CBM51b is somewhat unique among CBMs.

The remaining CBM51 subfamilies are uncharacterized. However, it is notable that the galactose-binding residues identified in GH95CBM51 are conserved among these subfamilies (data not shown), suggesting that they may be related to CBM51b by an ability to bind galactose.

**Conclusion**—The results presented here confirm that the NPCBMs do indeed constitute a new family of CBMs. Bioinformatic analysis of this family showed that its members form a distinct cluster in the phylogenetic tree. The remaining CBM51 subfamilies are uncharacterized.

**References**

1. Boraston, A. B., Bolam, D. N., Gilbert, H. J., and Davies, G. J. (2004) *Biochem. J.* **382**, 769–782
2. Bolam, D. N., Ciruela, A., Mcqueen-Mason, S., Simpson, P., Williamson, M. P., Rixon, J. E., Boraston, A., Hazlewood, G. P., and Gilbert, H. J. (1998) *Biochem. J.* **331**, 775–781
3. Ficko-Blean, E., and Boraston, A. B. (2006) *J. Biol. Chem.* **281**, 37748–37757
4. Boraston, A. B., Ficko-Blean, E., and Healey, M. (2007) *Biochemistry* **46**, 11352–11360
5. Boraston, A. B., Wang, D., and Burke, R. D. (2006) *J. Biol. Chem.* **281**, 35263–35271
6. Lammerts van Bueren, A., Higgins, M., Wang, D., Burke, R. D., and Boraston, A. B. (2007) *Nat. Struct. Mol. Biol.* **14**, 76–84
7. Rood, J. I. (1999) *Annu. Rev. Microbiol.* **53**, 333–360
8. Petit, L., Gilbert, M., and Popoff, M. R. (1999) *Trends Microbiol.* **7**, 104–110
9. Shimizu, T., Ohtani, K., Hirakawa, H., Ohshima, K., Yamashita, A., Shiba, T., Ogasawara, N., Hattori, M., Kuhara, S., and Hayashi, H. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99**, 996–1001
10. Myers, G. S., Rasko, D. A., Cheung, J. K., Ravel, J., Seshadri, R., DeBoy, R. T., Ren, Q., Varga, J., Awad, M. M., Brinkac, L. M., Daugherty, S. C., Haft, D. H., Dodson, R. J., Madupu, R., Nelson, W. C., Rosovitz, M. J., Sullivan, S. A., Khouri, H., Dimitrov, G. I., Watkins, K. L., Mulligan, S., Benton, J., Cadwell, D., Fisher, D. J., Atkins, H. S., Hiscox, T., Jost, B. H., Billington, S. J., Songer, J. G., McClane, B. A., Titball, R. W., Rood, J. I., Melville, S. B., and Paulsen, I. T. (2006) *Genome Res.* **16**, 1031–1040
11. Chitayat, S., Gregg, K., Adams, J. J., Ficko-Blean, E., Bayer, E. A., Boraston, A. B., and Smith, S. P. (2008) *J. Mol. Biol.* **375**, 20–28
12. Rigden, D. J. (2005) *FEBS Lett.* **579**, 5466–5472
13. Anderson, K. M., Ashida, H., Maskos, K., Dell, A., Li, S. C., and Li, Y.-T. (2005) *J. Biol. Chem.* **280**, 7720–7728
14. Wiseman, T., Williston, S., Brandts, J. F., and Lin, L. N. (1989) *Anal. Biochem.* **179**, 131–137
15. Pfugrath, J. W. (1999) *Acta Crystallogr*. Sect. D Biol. Crystallogr. **55**, 1718–1725
16. Evans, G., and Bricogne, G. (2002) *Acta Crystallogr*. Sect. D Biol. Crystallogr. **58**, 976–991
17. Cowtan, K. D., and Zhang, K. Y. (1999) *Prog. Biophys. Mol. Biol.* **72**, 245–270
18. Peraakis, A., Morris, R., and Lamzin, V. S. (1999) *Nat. Struct. Biol.* **6**, 458–463
19. Emsley, P., and Cowtan, K. (2004) *Acta Crystallogr*. Sect. D Biol. Crystallogr. **60**, 2126–2132
20. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) *Acta Crystallogr*. Sect. D Biol. Crystallogr. **53**, 240–255
21. Vagin, A., and Teplyakov, A. (2000) *Acta Crystallogr*. Sect. D Biol. Crystallogr. **56**, 1622–1624
22. Brunger, A. T. (1992) *Nature* **355**, 472–475
23. Atltschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402
24. Zdobnov, E. M., and Apweiler, R. (2001) *Bioinformatics* (Oxf.) **17**, 847–848
25. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22**, 4673–4680
26. Choi, I. H., Jung, H. Y., Kim, H. S., and Cho, H. G. (2000) *Bioinformatics* (Oxf.) **16**, 1056–1058
27. Krissinel, E., and Henrick, K. (2004) *Acta Crystallogr*. Sect. D Biol. Crystallogr. **60**, 2256–2268

**Structural and Functional Analysis of the CBM51 Family**

**Acknowledgments**—We are grateful to Cores D and H of the Consortium for Functional Glycomics for providing the blood group A and B carbohydrates used in the isothermal titration calorimetry experiments and performing the glycan array experiments, respectively. We thank Professor Yu-Teh Li for providing the C. perfringens ATCC 10543 genomic DNA.
28. Walser, P. J., Haebel, P. W., Kunzler, M., Sargent, D., Kues, U., Aebi, M., and Ban, N. (2004) Structure 12, 689–702
29. Cao, S., Lou, Z., Tan, M., Chen, Y., Liu, Y., Zhang, Z., Zhang, X. C., Jiang, X., Li, X., and Rao, Z. (2007) J. Virol. 81, 5949–5957
30. Kulkarni, K. A., Katiyar, S., Surolia, A., Vijayan, M., and Suguna, K. (2007) Proteins 68, 762–769
31. Boraston, A. B., Nurizzo, D., Notenboom, V., Ducros, V., Rose, D. R., Kilburn, D. G., and Davies, G. I. (2002) J. Mol. Biol. 319, 1143–1156
32. Henshaw, J., Horne, A., Lammerts van Bueren, A., Money, V. A., Bolam, D. N., Czjzek, M., Ekberg, N. A., Weiner, R. M., Hutcheson, S. W., Davies, G. J., Boraston, A. B., and Gilbert, H. J. (2006) J. Biol. Chem. 281, 17099–17107
33. Simpson, P. J., Xie, H., Bolam, D. N., Gilbert, H. J., and Williamson, M. P. (2000) J. Biol. Chem. 275, 41137–41142
34. Lammerts van Bueren, A., and Boraston, A. B. (2004) J. Mol. Biol. 340, 869–879
35. Read, R. J. (1986) Acta Crystallogr. Sect. A 42, 140–149