Carbohydrate Recognition by an Architecturally Complex α-*N*-Acetylglucosaminidase from *Clostridium perfringens*

Elizabeth Ficko-Blean[1], Christopher P. Stuart[1], Michael D. Suits[1], Melissa Cid[1], Matthew Tessier[2], Robert J. Woods[2,3], Alisdair B. Boraston[1]*

[1] Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia, Canada, [2] Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia, United States of America, [3] School of Chemistry, National University of Ireland, Galway, Ireland

Abstract

CpGH89 is a large multimodular enzyme produced by the human and animal pathogen *Clostridium perfringens*. The catalytic activity of this *exo*-α-D-*N*-acetylglucosaminidase is directed towards a rare carbohydrate motif, *N*-acetyl-β-D-glucosamine-α-1,4-D-galactose, which is displayed on the class III mucins deep within the gastric mucosa. In addition to the family 89 glycoside hydrolase catalytic module this enzyme has six modules that share sequence similarity to the family 32 carbohydrate-binding modules (CBM32s), suggesting the enzyme has considerable capacity to adhere to carbohydrates. Here we suggest that two of the modules, CBM32-1 and CBM32-6, are not functional as carbohydrate-binding modules (CBMs) and demonstrate that three of the CBMs, CBM32-3, CBM32-4, and CBM32-5, are indeed capable of binding carbohydrates. CBM32-3 and CBM32-4 have a novel binding specificity for *N*-acetyl-β-D-glucosamine-α-1,4-D-galactose, which thus complements the specificity of the catalytic module. The X-ray crystal structure of CBM32-4 in complex with this disaccharide reveals a mode of recognition that is based primarily on accommodation of the unique bent shape of this sugar. In contrast, as revealed by a series of X-ray crystal structures and quantitative binding studies, CBM32-5 displays the structural and functional features of galactose binding that is commonly associated with CBM family 32. The functional CBM32s that CpGH89 contains suggest the possibility for multivalent binding events and the partitioning of this enzyme to highly specific regions within the gastrointestinal tract.

Citation: Ficko-Blean E, Stuart CP, Suits MD, Cid M, Tessier M, et al. (2012) Carbohydrate Recognition by an Architecturally Complex α-*N*-Acetylglucosaminidase from *Clostridium perfringens*. PLoS ONE 7(3): e33524. doi:10.1371/journal.pone.0033524

Editor: Maria Gasset, Consejo Superior de Investigaciones Cientificas, Spain

Received November 8, 2011; **Accepted** February 14, 2012; **Published** March 27, 2012

Copyright: © 2012 Ficko-Blean et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by a Canadian Institutes of Health Research grant (FRN 68913). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: boraston@uvic.ca

Introduction

Mucins are heavily O-glycosylated glycoproteins that act to protect the epithelia from harmful bacteria by forming a biophysical barrier to infection as well as supporting innate and adaptive immunity [1]. A heavily hydrated and highly viscous protective mucosal layer can be found lining the surface of the major entry points to our body, including the eyes, the nasopharynx, the genito-urinary tract and the gastrointestinal tract. Within the gastrointestinal tract the mucin layer can vary from 700 μm deep in the stomach to 150–300 μm deep in the small intestine [2]. Pathogens of the gastrointestinal tract, such as *Clostridium perfringens*, must find ways to subvert or somehow challenge this protective mucosal barrier in order to set up infection.

C. perfringens' niche environment is in the gut of animals, including humans, where it may reside harmlessly; however, infection with a pathogenic strain can cause gastroenteritis and, in serious cases, substantial intestinal tissue destruction associated with necrotic enteritis. Among the enzymes that *C. perfringens* employs to cope with the mucosal surface are the glycoside hydrolases, which have varying catalytic specificities that reflect the diversity in host glycans; these include, but are not limited to, neuraminidases (GH33)[3,4], exo- and endo-β-*N*-acetylglucosaminidases (GH84 and GH85)[5,6,7], an endo-α-*N*-acetylgalactosaminidase (GH101)[8,9], as well as CpGH89, which is an exo-α-*N*-acetylglucosaminidase [10,11]. Due to the significant genome content of genes encoding carbohydrate-active enzymes with known or suspected specificity for complex glycans, such as those found on the mucosal surface, it has been postulated that these enzymes play an important role during colonization and/or infection. Indeed, enzymatic preparations of *C. perfringens*, in combination with mild acid hydrolysis, have previously been used to help partially "untangle" the complex carbohydrate surface lining the gut supporting the concept that the structure of gastrointestinal mucosa can be influenced by these bacterial factors [12].

Within the gastric mucosa there are two types of mucous cells, surface mucous cells and the deeper gland mucous cells, producing two different mucins which combine together to form a stratified surface mucous layer [13]. Class III mucins are produced normally by the gastric gland mucous cells, duodenal Brunner's gland mucous cells, and the mucous cells of the accessory glands of pancreaticobiliary tract but also in certain tissues exhibiting gastric metaplasia or adenocarcinoma [14–23]. The class III mucins, discharged by gland mucous cells in the gastric pits [13], are somewhat distinct in that they are specifically decorated with peripheral α-GlcNAc (α-*N*-acetyl-D-glucosamine) residues forming

GlcNAc-α-1,4-Gal-β-R (*N*-acetyl-β-D-glucosamine-α-1,4-D-galactose) motifs [19,22,24]. The biological relevance of this carbohydrate motif is at present not clear; however, terminal α-linked GlcNAc has been implicated as a host defense mechanism against colonization of the gastric mucosa by *Helicobacter pylori* [25] by blocking production of CGL (cholesteryl-α-D-glucopyranoside), an important component of this bacterium's cell wall.

C. perfringens is unusual in its ability to process the GlcNAc-α-1,4-Gal motifs found in class III mucin. CpGH89 (EC 3.2.1.50, CPF_0859), also referred to as AgnC [11], is a family 89 α-*N*-acetyl-D-glucosaminidase that has been shown to specifically release terminal α-linked GlcNAc from the disaccharide GlcNAc-α-1,4-Gal and demonstrated to liberate GlcNAc from crude class III porcine gastric mucin [10,11]. Using a *cpgh89* mutant of *C. perfringens* the activity of CpGH89 has been linked to the ability of *C. perfringens* to grow on mucin bearing this rare carbohydrate motif [11].

Two remarkable features of CpGH89 are its overall size (2095 amino acids) and its extensive multimodularity. Overall, the enzyme comprises a glycoside hydrolase family 89 (GH89) catalytic module, four FIVAR (found in various molecular architectures) modules, an unknown module, a C-terminal fibronectin type III-like (FN3-like) module, and six putative carbohydrate-binding modules (CBMs) (Figure 1). CBMs are generally defined as non-catalytic modules that bind carbohydrates and are found within the modular architectures of carbohydrate-active enzymes [26], thus distinguishing these modules from lectins and carbohydrate-specific antibodies. CBMs are presently classified into over 60 amino acid sequenced based families; the CBMs from CpGH89 all belong to CBM family 32, which is one of the most diverse CBM families [7].

Based on truncation studies of the enzyme and structural analyses of the N-terminal modules, the catalytic activity of the enzyme allowing it to release GlcNAc from class III mucin is attributed to its GH89 module [10,11]. Similar truncation studies that focused solely on CBM32s 2 to 6 revealed one or more of these CBMs to be able to bind mucin [11]. Notably, constructs of CpGH89 lacking the three most C-terminal CBMs had reduced activity on mucin suggesting an important role for the CBMs in substrate recognition. Thus, CpGH89 possesses a complex multimodular architecture where the composite modules function together to efficiently act on components of mucin. Though it is clear that the CBMs are able to bind mucin what remains unknown is what carbohydrate motifs displayed on mucin, particularly the unique GlcNAc-α-1,4-Gal motif, may be recognized by the CBM32s and what the molecular bases of these interactions are. Here we address these questions through structural and functional analyses of the CBMs from CpGH89. Overall, these studies reveal the specificity of three of CBM32s and, through X-ray crystal structures, how two of the CBMs accommodate their ligands, which includes the first GlcNAc-α-1,4-Gal binding specificity for a protein other than an antibody.

Results and Discussion

Analysis of a galactose binding CBM

Of the six putative CBM32s in CpGH89 CBM32-5, the fifth CBM, has the highest similarity with modules known to have carbohydrate-binding function (~43% amino acid sequence identity with the CBM32 from the large sialidase NanJ, also from *C. perfringens*). Furthermore, the strict conservation of residues involved in galactose recognition suggested that CBM32-5 belongs to the galactose binding group of family 32 CBMs [7,27,28]. CBM32-5 was initially screened for carbohydrate binding on glycan microarrays. Binding was generally quite weak; however, two galactose terminating N-glycans, one tri-antennary and the other tetra-antennary, gave significant binding signal (Figure 2A). Likewise, two glycans terminating with GalNAc, one α-1,4-linked and the other β-1,3-linked, also gave good signals. Though this did not conclusively single out a single carbohydrate ligand it is generally consistent with predictions of galactose specificity based on amino acid sequence similarity. This suggested binding to terminal galactose and GalNAc residues, which was used as a guide to quantitatively assess binding to carbohydrate ligands.

The addition of galactose or GalNAc to CBM32-5 perturbed the UV absorption of this protein in a manner consistent with the involvement of tyrosine residues in carbohydrate binding [29](Figure 2B). This signal was used in a quantitative manner to assess binding to a variety of carbohydrate ligands (Figure 2C and Table 1). The association constants of CBM32-5 binding to ligands containing galactose or GalNAc were in the range of 2–5×10^{3} M^{-1} (Figure 2B, 2C and Table 1), and thus quite weak, but of the same magnitude observed for other family 32 CBMs [3,27,30–32]. The CBM displayed little to no preference for either galactose or GalNAc and did not appear to significantly favor common disaccharide motifs that terminate in galactose or GalNAc over the monosaccharides (Table 1).

The structural basis for what appears to be a general selectivity for terminal galactose residues was examined by determining the X-ray crystal structure of CBM32-5 in complex with carbohydrate. The 1.55 Å resolution structure of the CBM binding galactose revealed the β-sandwich fold with structural metal ion, in this case modeled as a Ca^{2+}, which is common to the family (Figure 3A). The galactose residue was well-ordered in the crystal structure providing clear electron density (Figure 3B). The site accommodating this carbohydrate is a shallow cleft marked by two solvent exposed aromatic side chains, F1483 and Y1395 (Figure 3C), which is present in the loops at the edges of the β-sandwich (Figure 3A). The C6-OH group of galactose fits into a corner of the binding site made up by F1483 and Y1395, whose aromatic rings are at nearly right angles to one another (Figure 3C and 3D). A series of hydrogen bonds involve the side chains of four amino acids in the carbohydrate-binding site (Figure 3D). With the exception of E1376, which makes hydrogen bonds with the C3 hydroxyl group of galactose, all of the interactions are highly conserved with other known galactose binding CBMs (Figure 3E). Indeed, the interactions made by the five residues H1392, Y1395,

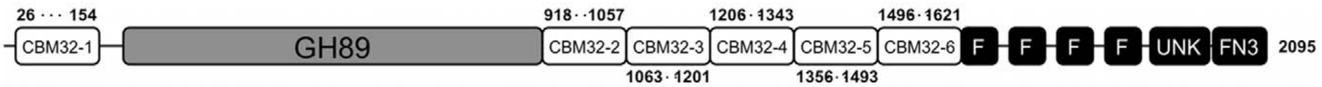


Figure 1. Schematic representation of the modular structure of CpGH89. CBM32 denotes family 32 carbohydrate-binding modules, GH89 represents the family 89 catalytic module, F denotes the predicted FIVAR (Found In Various Architectural Regions) modules and FN3 refers to a fibronectin type III domain. Modular boundaries used in this study are given above and below the schematic.
doi:10.1371/journal.pone.0033524.g001

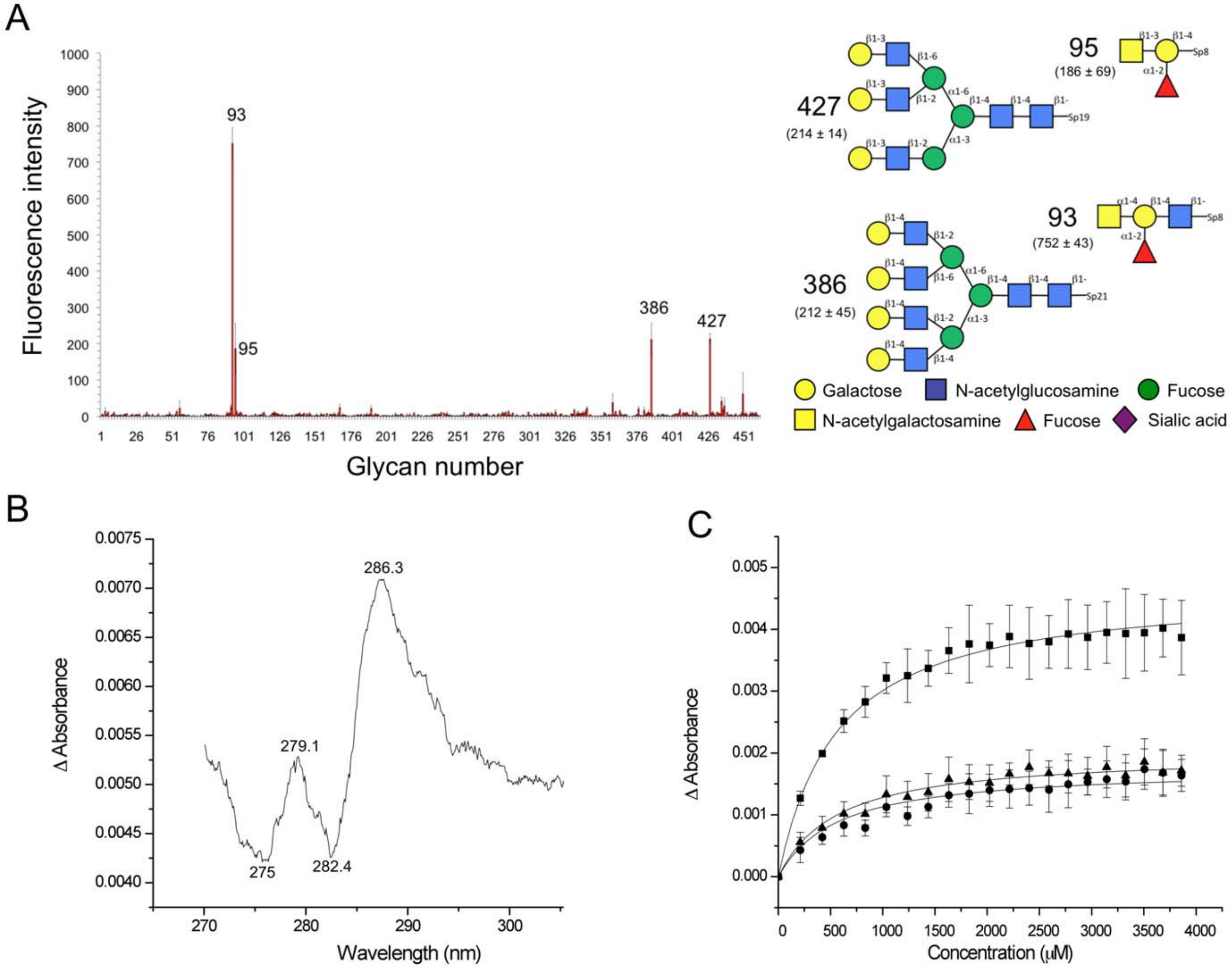


Figure 2. Analysis of carbohydrate binding by CBM32-5. (A) Glycan microarray analysis of carbohydrate binding by CBM32-5. The carbohydrates giving the most significant signals are numbered and their structures are shown schematically to the immediate right. The fluorescence intensities measure for the glycans are shown with the structural schematics. The reported error represents the standard error of the mean for quadruplicate measurements. A legend to symbols representing specific monosaccharides is provided. (B) UV difference spectrum upon CBM32-5 binding to excess galactose. Wavelengths of peaks and troughs are labeled. (C) Binding isotherm of CBM32-5 binding to galactose generated by UV difference titrations. The squares, triangles and circles represent the values for the peak to trough height differences for the wavelength pairs of 287.2/282.4 nm, 279.1/275.3 nm, and 279.1/282.4 nm, respectively. The error bars represent the standard deviations from triplicate independent titrations. The solid lines show the fits to a one-site binding model.
doi:10.1371/journal.pone.0033524.g002

Table 1. Affinity of CBM32-5 for carbohydrates determined at 20°C in 20 mM Tris HCl, pH 8.0.

Carbohydrate	K_a (M^{-1})
D-galactose	1.69 (±0.05)×10^3
D-GalNAc	5.01 (±0.64)×10^3
Lactose (Gal-β-1,4-Glc)	2.40 (±0.58)×10^3
Gal-β-1,3-GalNAc	3.77 (±0.34)×10^3
GalNAc-β-1,3-Gal[a]	3.40 (±0.10)×10^3

[a]the binding constant for GalNAc-β-1,3-Gal was determined by ITC; the remaining binding constants were determined by UV difference titrations.
doi:10.1371/journal.pone.0033524.t001

R1423, N1428, and F1483 make up the canonical galactose-binding motif in the family 32 CBMs [7,27,32]. CBM32-5, therefore, possesses a galactose-binding site; however, it is also capable of binding GalNAc equally well. Furthermore, the analysis of the CBM32 from NagJ, indicated that the recognition of longer glycans by CBM32s can involve additional subsites [27]. The structures of CBM32-5 in complex with other potentially biologically relevant ligands, GalNAc, the Tn-antigen, and GalNAc-β-1,3-Gal (Figure 4A, 4B and 4C) show the recognition of terminal GalNAc residues to be identical to that of galactose, with the addition of a water mediated hydrogen bond involving the acetamido group of the carbohydrate and the backbone nitrogens of K1427 and N1428 (Figure 4D). This limited additional interaction appears to provide little to no favorable energy to binding. Likewise, the galactose of the GalNAc-β-1,3-Gal extended away from the protein surface and made no interactions with the protein, which is consistent with the lack of improved binding for this disaccharide over GalNAc. The same observation was made for the serinyl group of the Tn-antigen, even though the serine is α-linked to GalNAc. Modeling other common α-linked carbohydrates, such as Gal-α-1,3-Gal, based on the Tn-antigen complex suggested that these additional residues also extend out into solvent with no capacity to make additional interactions with the protein (not shown).

The crystallography results suggest that CBM32-5 is relatively promiscuous in that it requires only a terminal galactose or GalNAc residue with little preference for the sugar that precedes it. The glycan microarray results, however, suggested a strong interaction with a unique carbohydrate, GalNAc-α-1,4(Fuc-α-1,2)-Gal-β-1,4-GlcNAc. This interaction was reproducible on glycan microarrays, even when using CBM that was directly labeled by chemically coupling the fluorophore to primary amines on the CBM (not shown). To our knowledge, this glycan has not been identified in any mammalian tissues; however, this synthetic carbohydrate was clearly the top ligand from the array analysis suggesting that an analysis of the interaction of CBM32-5 with this carbohydrate may provide insight into the recognition of more complex but as yet unstudied glycans. A molecular dynamics approach was used to study the potential interaction of GalNAc-α-1,4(Fuc-α-1,2)-Gal-β-1,4-GlcNAc-OMe with CBM32-5. The resulting analysis gave an ensemble of ten structures with each structure representing a group of similar, energy-minimized structures (Figure 4E). Overall, the carbohydrate in the ten structures adopts an array of potential conformations, though the terminal GalNAc residue and the preceding Gal residue are somewhat constrained in their positions. A representative of the lowest energy group of models shows the carbohydrate to adopt a conformation that, by virtue of the bent conformation imparted by the α-1,4-linkage between the GalNAc and Gal, bends around Y1395 and allows the reducing-end portion of the glycan to rest against the protein surface with only a very small number of additional hydrogen bonds made (Figure 4F). Free energy

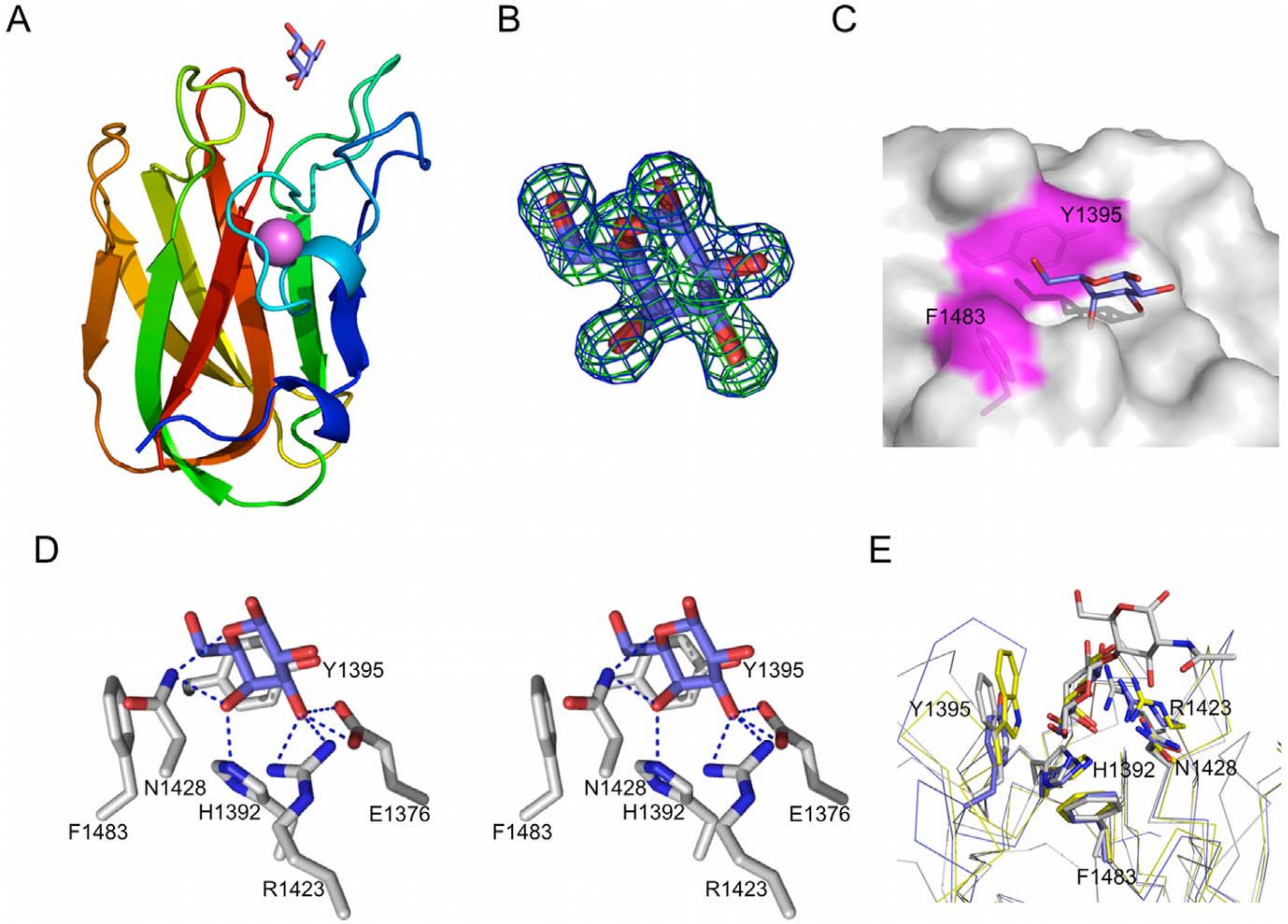


Figure 3. Structural analysis of the interaction between CBM32-5 and galactose. (A) A cartoon representation of the structure of CBM32-5 bound to galactose (blue sticks) determined by X-ray crystallography to 1.55 Å resolution. The bound calcium atom is shown as a pink sphere. (B) Electron density for galactose within the binding site of CBM32-5. Electron density maps are maximum-likelihood/σ_A [59] -weighted $2F_{obs}$-F_{calc} contoured at 1 σ (both maps at 0.45 $e^{-}/Å^{3}$) produced by refinements prior to modeling the sugar (green) and with the monosaccharide included (blue). (C) Surface representation of the CBM32-5 binding site with the bound galactose shown as blue sticks. The aromatic amino acids providing the hydrophobic binding platform are shown as sticks and labeled while the surface they contribute to the active site is coloured magenta. (D) Divergent stereo view of the key interactions between the binding site of CBM32-5 and galactose. Hydrogen bonds are shown as dashed black lines. (E) Comparison of the binding site of CBM32-5 (blue) with the CBM32 from *C. perfringens* GH84C (grey; PDB code 2J1E) and the CBM32 from *C. perfringens* NanJ (yellow; PDB code 2V72) reveals the canonical galactose-binding site. Conserved amino acid side chains and bound carbohydrates are shown as sticks.

doi:10.1371/journal.pone.0033524.g003

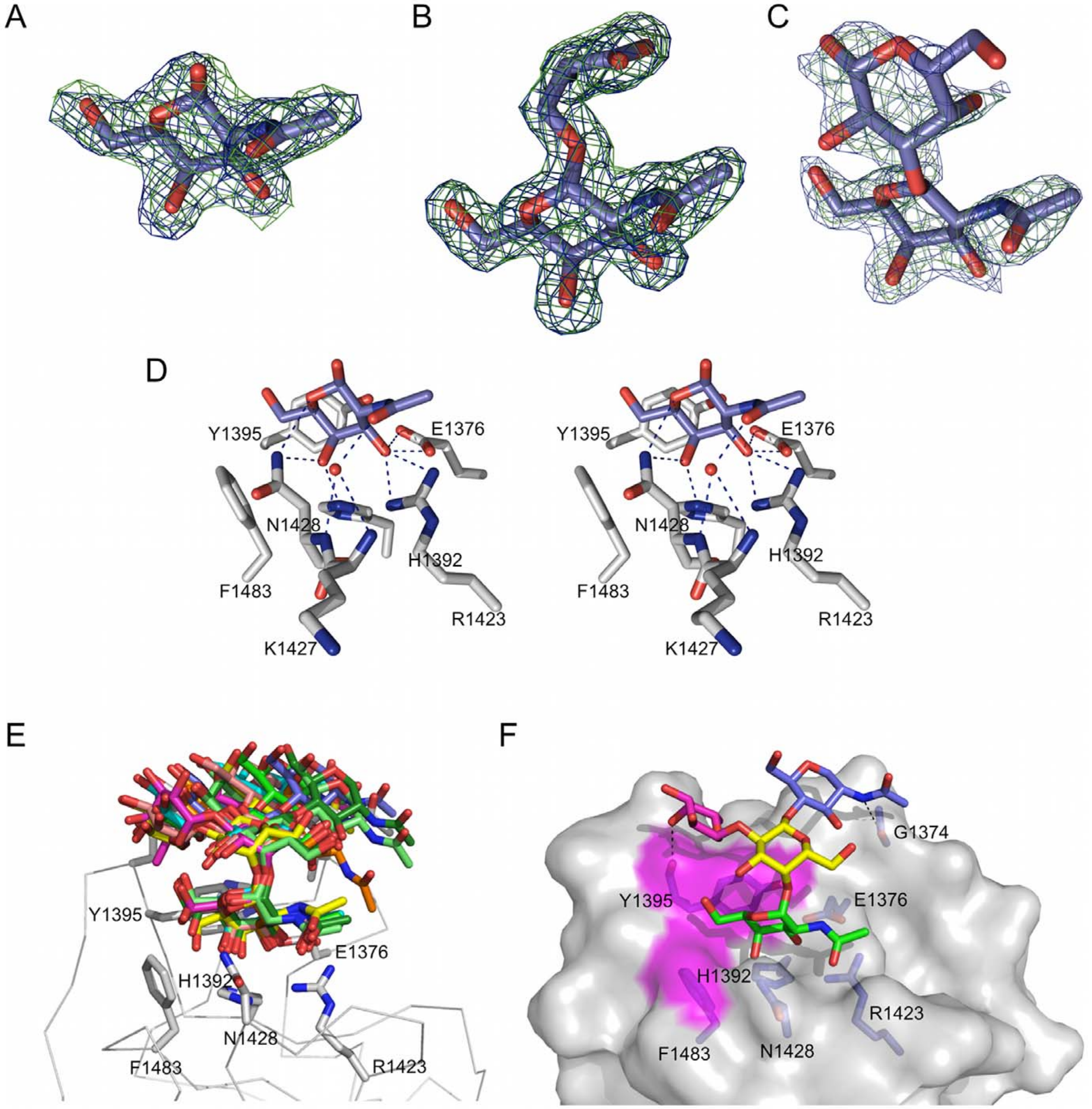


Figure 4. Structural analysis of CBM32-5 with additional carbohydrates. (A) Electron density for GalNAc shown as maximum-likelihood/σ_A [59] -weighted 2F_{obs}-F_{calc} maps contoured at 1 σ (both maps at 0.31 $e^-/Å^3$) produced by refinements prior to modeling the sugar (green) and with the sugar included (blue). (B) Electron density for serinyl-Tn antigen shown as maximum-likelihood/σ_A [59] -weighted 2F_{obs}-F_{calc} maps contoured at 1 σ (both maps at 0.45 $e^-/Å^3$) produced by refinements prior to modeling the sugar (green) and with the sugar included (blue). (C) Electron density for GalNAc-β-1,3-galactose shown as maximum-likelihood/σ_A [59] -weighted 2F_{obs}-F_{calc} maps contoured at 0.8 σ (both maps at 0.34 $e^-/Å^3$) produced by refinements prior to modeling the sugar (green) and with the sugar included (blue). (D) Divergent stereo view of the key interactions between the binding site of CBM32-5 and GalNAc. This also represents the mode of interaction between the CBM and the serinyl-Tn antigen and GalNAc-β-1,3-galactose, which all have identical hydrogen bonding patters. Hydrogen bonds are shown as dashed black lines. (E) Models of CBM32-5 in complex with GalNAc-α-1,4(Fuc-α-1,2)-Gal-β-1,4-GlcNAc-OMe produced by molecular dynamics simulations. An ensemble of ten energy minimized models is given with each model representing a group of energetically similar models. Relevant residues in the binding site are shown as grey sticks with the backbone of the protein shown as a Cα-ribbon. (F) A surface representation of the lowest energy model of CBM32-5 bound to the tetrasaccharide (GalNAc is shown in green, fucose in pink, galactose in yellow, and GlcNAc in blue). The surfaces contributed by Y1395 and F1483 are shown in magenta and additional hydrogen bonds made outside of the primary galactose binding site shown as dashed lines.
doi:10.1371/journal.pone.0033524.g004

decomposition shows the increased affinity of this ligand for CBM32-5 results from the increased van der Waals and non-polar solvation interactions that is imparted by the complementary interacting surface areas of this unique carbohydrate ligand and the CBM surface. This interaction is specifically enhanced by interactions between the fucosyl residue and residues Y1395 and N1396 of CBM32-5 (Figure S1). Though GalNAc-α-1,4(Fuc-α-1,2)-Gal-β-1,4-GlcNAc may not be a biologically relevant ligand for CBM32-5 its mode of interaction with this CBM suggests that other high affinity ligands, perhaps not represented on the carbohydrate microarrays, may be possible provided they adopt a conformation that maximizes the interacting surface areas.

Carbohydrate-binding modules with unique specificity

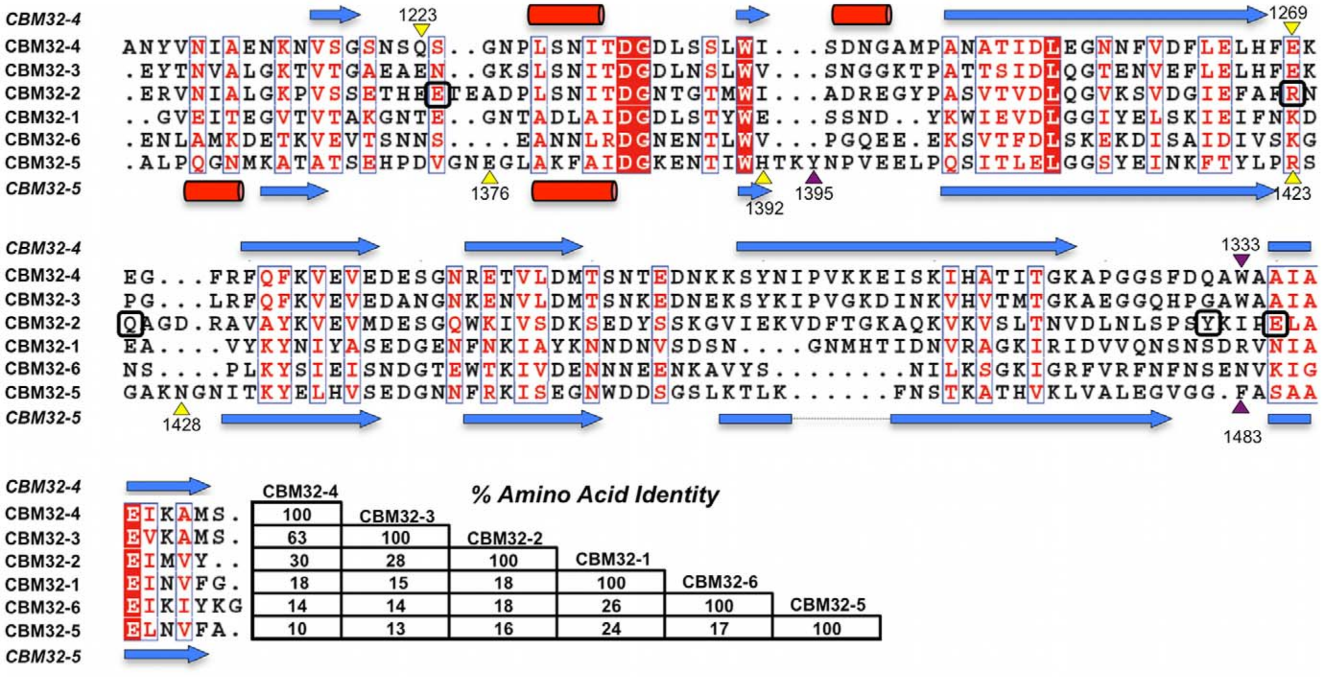
Though CpGH89 has at least one functional CBM its specificity (i.e. galactose and GalNAc) is clearly mismatched with the specificity of the catalytic module. Furthermore, this CBM is an outlier among the CpGH89 CBMs as it has higher amino acid sequence identity with CBMs from other enzymes than it does with the remaining CBMs from CpGH89. In contrast, CBM32-2, CBM32-3, and CBM32-4 form a distinct cluster in the phylogenetic analysis of the CBM32 family [7]. Indeed, CBM32-3 and CBM32-4 share 63% amino acid sequence identity and CBM32-2 has ~30% amino acid identity with these two CBMs (Figure 5). These putative CBMs have very low amino acid sequence identity with CBM32-5 and other known CBM32s suggesting they may represent a new functional class of CBM32s. Isolated CBM32-2, CBM32-3, and CBM32-4 were screened for binding on the glycan microarrays. CBM32-3 gave statistically meaningful binding (i.e. signal with standard errors of the mean that indicated significant binding above background) with the top hits terminating in GlcNAc-α-1,4-Gal (Figure 6A). Unfortunately, the results for CBM32-2 and CBM32-4 were inconclusive; however, the high amino acid sequence similarity between CBM32-3 and CBM32-4 suggested that both CBMs may have the same ligand, GlcNAc-α-1,4-Gal. Indeed, using ITC, the association constant of CBM32-4 for GlcNAc-α-1,4-Gal was determined to be $1.38\ (\pm 0.08) \times 10^4\ \mathrm{M}^{-1}$ thus showing this to be a relatively strong interaction for a family 32 CBM (Figure 6B). The titration of GlcNAc-α-1,4-Gal into CBM32-3 also produced a binding isotherm consistent with carbohydrate binding and the association constant was determined to be $2.64\ (\pm 0.64) \times 10^4\ \mathrm{M}^{-1}$ (Figure 6C). Thus, both CBM32-3 and CBM32-4 appear to have binding specificity for GlcNAc-α-1,4-Gal, which is complementary to the specificity of the catalytic module.

The ability of CBM32-3 and CBM32-4 to bind the GlcNAc-α-1,4-Gal is unique among non-catalytic carbohydrate binding proteins prompting the study of the molecular basis of this interaction. Of the two CBMs, crystals were only obtained of CBM32-4. The structure of seleno-methionine labeled CBM32-4 was determined by single anomalous dispersion to 1.55 Å resolution. This CBM adopts a β-sandwich fold with conserved structural metal ion, modeled as a calcium atom, which is similar to that of CBM32-5 (root mean square deviation of 1.9 Å over 112 matched Cα) (Figure 7A). CBM32-4 was co-crystallized with GlcNAc-α-1,4-Gal and this structure determined to 2.8 Å resolution (Figure 7B). Both molecules of CBM32-4 in the asymmetric unit had bound disaccharide as revealed by clear electron density for the sugar located in the loops at the edges of the β-sandwich core (Figure 7B, C, and D). CBM32-4 accommodates the disaccharide in a shallow depression; the sugar, with its bent conformation, lies on edge in the depression with the B-face of the galactose residue pushed up against the planar surface of the W1333 side chain. Though there are no aromatic residues present on the adjacent wall of the binding site, it is at roughly right angles to the plane of the W1333 side chain and thus well positioned to pack against the A-face of the GlcNAc residue. Markedly few hydrogen bonds are made between the sugar and binding site suggesting that binding and specificity for this disaccharide is driven primarily by hydrophobic and van der Waals forces and accommodation of the unique carbohydrate conformation. O1 of the galactose is completely exposed and oriented out into the bulk solvent illustrating how the CBM might tolerate extensions on the reducing end of the GlcNAc-α-1,4-Gal motif, which is consistent with binding to the glycan microarrays and to the recognition of the motif as it would naturally be displayed at the termini of glycans on mucin. The O3 and O4 groups on the terminal GlcNAc, though solvent exposed, lie very close to the protein surface. It is unclear whether modification to these could be tolerated by the CBM, thereby allowing it to recognize internal GlcNAc-α-1,4-Gal motifs, but the proximity to the protein surface and steric clashes that would likely ensue suggests that this is unlikely. The C6 hydroxyl group is buried in the base of the binding site and thus extension with additional sugar residues would not be tolerated.

CBM32-3 was recalcitrant to crystallization preventing structural analysis by X-ray crystallography and direct examination of its interaction with carbohydrate; however, the main residues involved in GlcNAc-α-1,4-Gal recognition by CBM32-4 are conserved in CBM32-3 (Figure 5). Taking further advantage of the high amino acid sequence identity of the two CBMs, a homology model of CBM32-3 was constructed; this revealed not only conservation of the primary binding site residues but also the majority of the residues lining the binding site (Figure 7F), indicating that the mode of carbohydrate recognition by CBM32-3 is likely extremely similar to that of CBM32-4.

To date, the structural analysis of family 32 CBMs found in carbohydrate-active enzymes has revealed two subtypes of CBMs within the family: the 'canonical' galactose binding CBM32s, such as CBM32-5, and the unique GlcNAc binding CBM32 as represented by the CBM from NagH, NagHCBM32-2 [31]. A comparison of the amino acids involved in ligand binding from CBM32-4 with the binding sites of both of these CBM32 subtypes shows them to have no similarities in carbohydrate recognition beyond the general placement of the active sites (Figure 7G and 7H). Thus, the GlcNAc-α-1,4-Gal binding CBMs, CBM32-3 and CBM32-4, represent a new mode of carbohydrate recognition by the CBM32s and continue to highlight the diversity within this family of CBMs.

Glycan microarray binding experiments with CBM32-2 were inconclusive, as were other low-throughput experiments to identify potential ligands, and attempts at crystallization did not yield crystals of sufficient quality for structure determination. To provide some insight into the potential capacity of this module to interact with carbohydrate a homology model based on the structure of CBM32-4 was constructed. Though the residues in CBM32-4 that impart carbohydrate binding function are not conserved with CBM32-2 (Figure 5) the model reveals a pocket in the protein surface located in loops that usually contain the binding sites of CBM32s (Figure 8A and 8B). This pocket contains a solvent exposed aromatic amino acid, Y1046, and a series of exposed planar polar amino acid side chains (Figure 8B). These features are generally consistent with the properties of carbohydrate binding sites in CBMs, suggesting that this module is indeed capable of recognizing an as yet unidentified sugar.


CBM32-4
1223
1269
CBM32-4 ANYVNIAENKNVSGSNSQS..GNPLSNITDGDLSSLWI...SDNGAMPANATIDLEGNNFVDFLELHFEK
CBM32-3 .EYTNVALGKTVTGAEAEN..GKSLSNITDGDLNSLWV...SNGGKTPATTSIDLQGTENVEFLELHFEK
CBM32-2 .ERVNIALGKPVSSETHEETEADPLSNITDGNTGTMWI...ADREGYPASVTVDLQGVKSVDGIEFARRN
CBM32-1 ..GVEITEGVTVTAKGNTE..GNTADLAIDGDLSTYWE...SSND..YKWIEVDLGGIYELSKIEIFNKD
CBM32-6 .ENLAMKDETKVEVTSNNS....EANNLRDGNENTLWV...PGQEE.EKSVTFDLSKEKDISAIDIVSKG
CBM32-5 .ALPQGNMKATATSEHPDVGNEGLAKFAIDGKENTIWHTKYNPVEELPQSITLELGGSYEINKFTYLPRS
CBM32-5
1376
1392
1395
1423
CBM32-4
1333
CBM32-4 EG...FRFQFKVEVEDESGNRETVLDMTSNTEDNKKSYNIPVKKEISKIHATITGKAPGGSFDQAWAAIA
CBM32-3 PG...LRFQFKVEVEDANGNKENVLDMTSNKEDNEKSYKIPVGKDINKVHVTMTGKAEGGQHPGAWAAIA
CBM32-2 QAGD.RAVAYKVEVMDESGQWKIVSDKSEDYSSKGVIEKVDFTGKAQKVKVSLTNVDLNLSPSYKIEELA
CBM32-1 EA....VYKYNIYASEDGENFNKIAYKNNDNVSDSN....GNMHTIDNVRAGKIRIDVVQNSNSDRVNIA
CBM32-6 NS....PLKYSIEISNDGTEWTKIVDENNNEENKAVYS........NILKSGKIGRFVRFNFNSENVKIG
CBM32-5 GAKNGNITKYELHVSEDGNNFRKISEGNWDDSGSLKTLK......FNSTKATHVKLVALEGVGG.FASAA
CBM32-5
1428
1483
CBM32-4
CBM32-4 EIKAMS.
CBM32-3 EVKAMS.
CBM32-2 EIMVY..
CBM32-1 EINVFG.
CBM32-6 EIKIYKG
CBM32-5 ELNVFA.
CBM32-5
% Amino Acid Identity

	CBM32-4	CBM32-3	CBM32-2	CBM32-1	CBM32-6	CBM32-5
CBM32-4	100					
CBM32-3	63	100				
CBM32-2	30	28	100			
CBM32-1	18	15	18	100		
CBM32-6	14	14	18	26	100	
CBM32-5	10	13	16	24	17	100

Figure 5. Amino acid sequence comparison of the CBM32 modules from CpGH89. The secondary structure is shown above (CBM32-4) and below (CBM32-5) with arrows representing β-strands and cylinders α-helices. The purple and yellow triangles above and below the sequences indicate the aromatic and hydrogen bonding residues, respectively, that are involved in carbohydrate binding by CBM32-4 (top) and CBM32-5 (bottom). Numbers with the triangles indicate the residue number. Residues in CBM32-2 that are highlighted by boxes are those present in the putative binding site of this module.
doi:10.1371/journal.pone.0033524.g005

CBM32-1 and CBM32-6 appear to lack carbohydrate-binding function

Despite the observation that CBM32-1 and CBM32-6 display only 26% amino acid identity (Figure 5) they cluster together in a phylogenetic analysis of CBM32 modules indicating that they are more closely related to one another than to other putative CBMs [7]. Qualitative UV difference scans on CBM32-1 and CBM32-6 did not suggest binding to any simple monosaccharides (galactose, GalNAc, mannose, sialic acid, GlcNAc or glucose). CBM32-1 was also screened on glycan microarrays but significant binding was not detected. The structure of CBM32-6 was determined to 1.55 Å resolution using SAD and seleno-methionine substituted protein (not shown). This structure compared with CBM32-1, previously determined as part of a construct including the catalytic module [10], gave a root mean square deviation of 1.8 Å over 119 Cα atoms. Neither CBM32-1 nor CBM32-6 have any exposed aromatics in the region of the protein known to contain the binding sites in CBM32 proteins (Figure 9). Furthermore, a more thorough analysis of the surface residues of CBM32-1 and CBM32-6 showed them both to lack features consistent with carbohydrate binding sites. This observation, along with the lack of experimental support for carbohydrate binding, suggest that CBM32-1 and CBM32-6 do not function as CBMs, which perhaps explains their somewhat outlying position in the phylogenetic analysis of CBM32 modules [7].

The modular diversity of CpGH89 and its implications

In order to colonize the gastrointestinal tract organisms must first infiltrate the mucosal surface. For example, the secreted mucosal surfaces of the colon are comprised of mainly Muc2, which forms both the thick outer mucous layer, that plays host to many commensal microbes, and the thin inner mucous layer that is impervious to bacteria [33,34]. GlcNAc-α-1,4-Gal is displayed by the deeper gastric-type mucosal class III mucins, Muc5Ac and Muc6 [19] and the catalytic activity of CpGH89 is directed at this specific carbohydrate structure. Furthermore, two of the CBMs in this enzyme, CBM32-3 and CBM32-4, have evolved binding specificity complementary to the catalytic specificity. In a manner consistent with the generally proposed role of CBMs [26], CBM32-3 and CBM32-4 likely direct the enzyme to the secreted class III mucins within the deep mucosa of the stomach and duodenum, and in doing so promote substrate degradation by the catalytic module. The presence of two CBMs with the same specificities indicate the potential for a multivalent interaction, thereby increasing the overall apparent affinity of the enzyme for regions that display clusters of the GlcNAc-α-1,4-Gal motif.

Of the six CBM32-like modules that CpGH89 possesses two do not appear to bind carbohydrate (their functions, if they have any, remain unknown), one has putative carbohydrate-binding function (CBM32-2), and the remaining three clearly have carbohydrate-binding function (CBM32-3, CBM32-4 and CBM32-5). The specificity of CBM32-5 appears to be primarily for terminal galactose and GalNAc residues and thus does not match the substrate preference of the catalytic module. Such mismatching between CBMs and their cognate catalytic modules is not unusual with *C. perfringens* glycoside hydrolases [3,27]. The biological reason for the presence of the mismatched CBMs remains speculative; however, it has been postulated that the presence of such CBMs may allow the enzyme to remain adhered to carbohydrate rich surfaces after the catalytic module has begun processing the substrate. For example, after hydrolysis of the GlcNAc-α-1,4-Gal substrate by the catalytic module of CpGH89 the remaining terminal sugar is a galactose residue and thus a

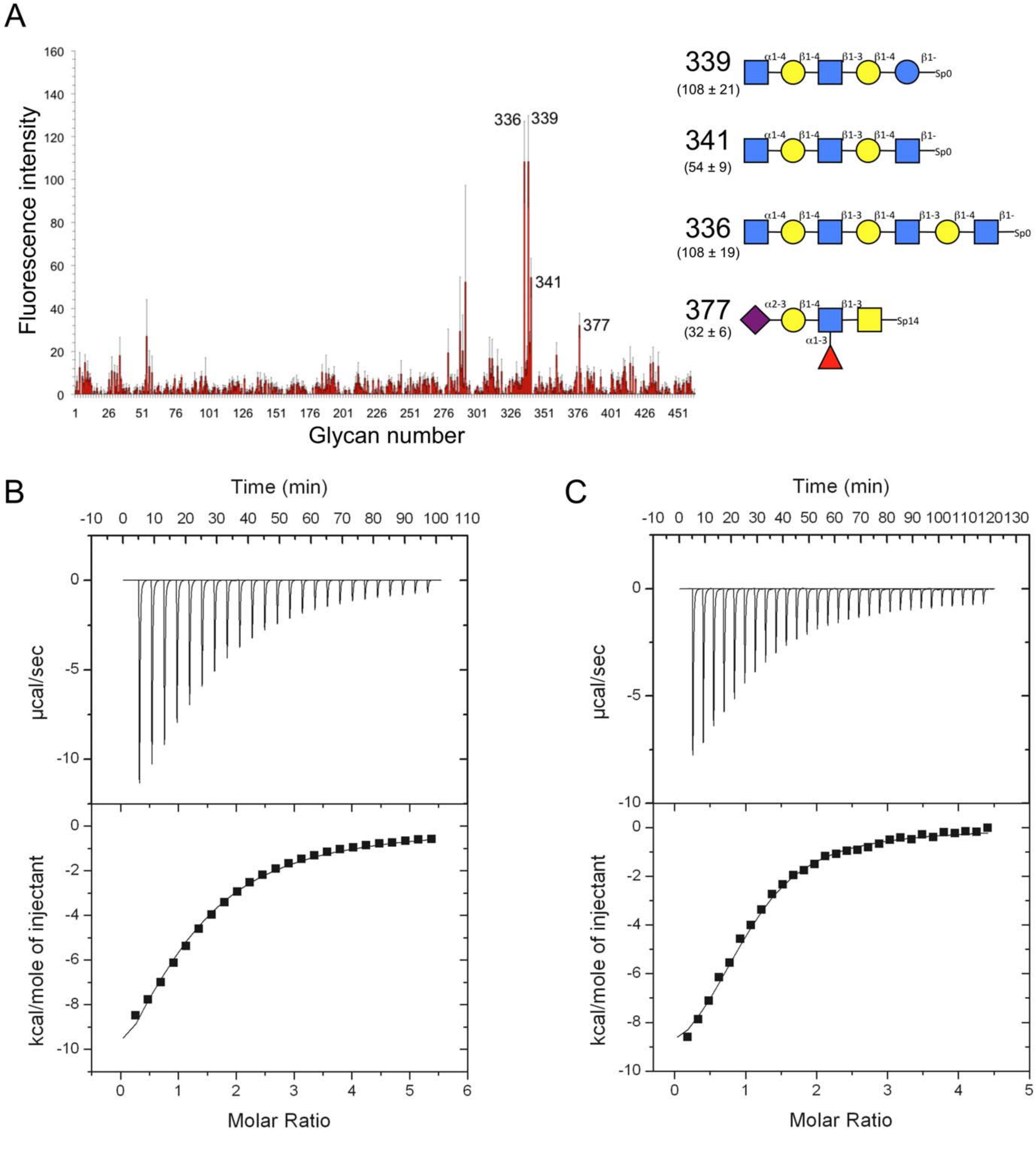


Figure 6. Analysis of carbohydrate binding by CBM32-3 and CBM32-4. (A) Glycan microarray analysis of carbohydrate binding by CBM32-3. The carbohydrates giving the most significant signals are numbered and their structures are shown schematically to the immediate right. The fluorescence intensities measured for the glycans are shown with the structural schematics. The reported error represents the standard error of the mean for quadruplicate measurements. The symbols representing specific monosaccharides are the same as those given for Figure 2A. (B) and (C) Representative isothermal calorimetry titrations of CBM32-4 and CBM32-3, respectively, binding to GlcNAc-α-1,4-Gal. Top portions of each panel show the raw power data while the bottom portions show the integrated and heat of dilution corrected data. Solid lines show the non-linear curve fits to a one site binding model. doi:10.1371/journal.pone.0033524.g006

potential ligand of CBM32-5. There then exists the potential for multivalent interactions involving heterogeneous clusters of ligands, such as combinations of the GlcNAc-α-1,4-Gal motif and terminal galactose and GalNAc residues. Alternatively, it has been hypothesized that the majority of the *C. perfringens* glycoside hydrolases, including CpGH89, are either covalently or non-covalently associated with the bacterial surface [6]. Thus, though the intrinsic affinity of a single CBM32-5 module for terminal

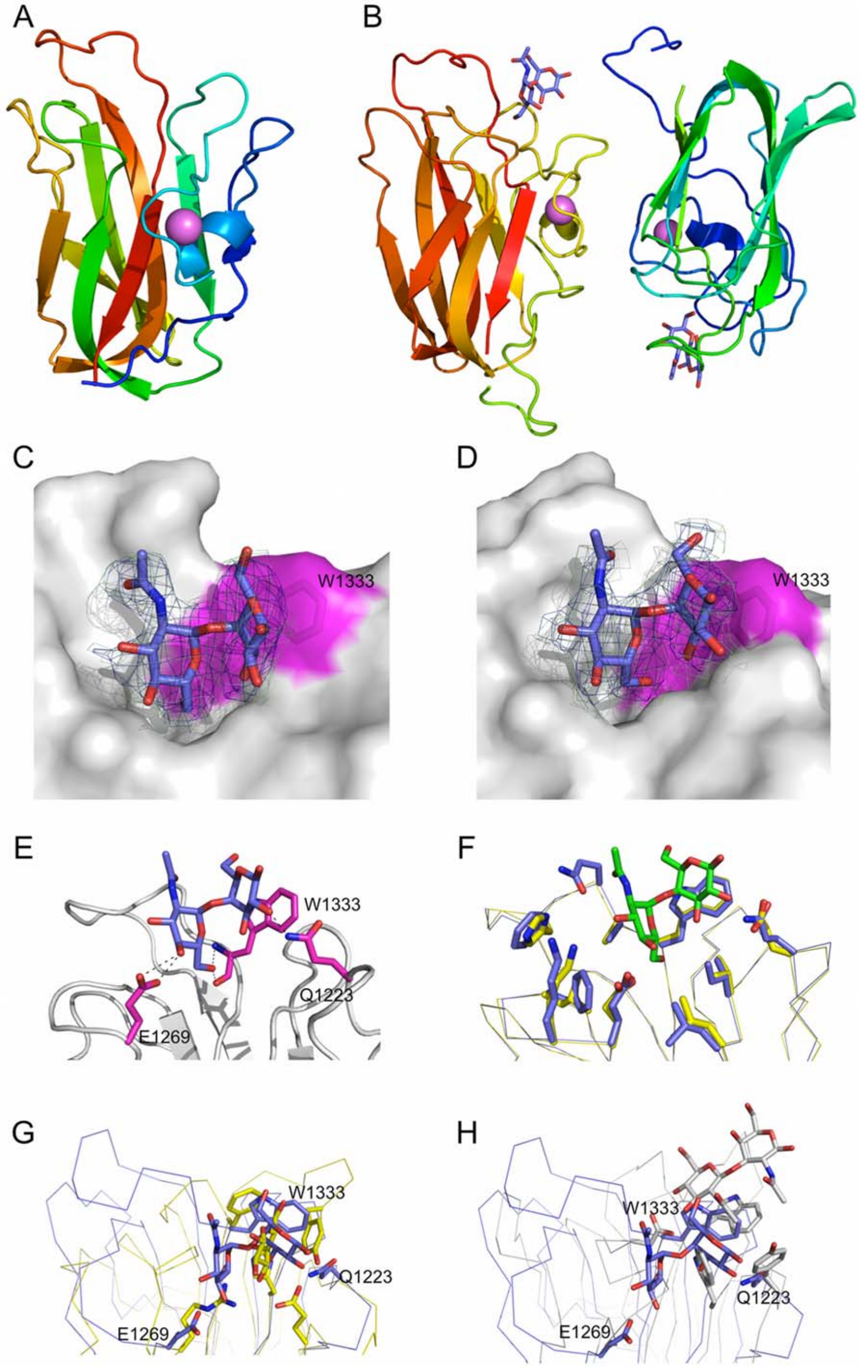


Figure 7. Structural analysis of the interaction between CBM32-4 and GlcNAc-α-1,4-Gal. (A) A cartoon representation of the structure of seleno-methionine labeled CBM32-4 determined by X-ray crystallography to 1.55 Å resolution. The bound calcium atom is shown as a pink sphere. (B) A cartoon representation of the dimer of CBM32-4 in complex with GlcNAc-α-1,4-Gal determined to 2.8 Å resolution. The bound calcium atom is shown as a pink sphere and the carbohydrate as blue sticks. (C) and (D) Electron density for GlcNAc-α-1,4-Gal in the binding sites of each monomer of CBM32-4 in the asymmetric unit (panel C shows molecule A and panel D shows molecule B). In both panels the electron density is shown as maximum-likelihood/σ_A [59] -weighted 2F_{obs}-F_{calc} maps contoured at 1 σ (all maps at 0.39 e^-/Å^3) produced by refinements prior to modeling the sugar (green) and with the sugar included (blue). Protein is shown as solvent accessible surface with the surface contributed by W1333 shown and labeled) colored magenta. (E) Key interactions between the binding site of CBM32-4 and the disaccharide as represented by monomer A in the asymmetric unit. (F) An overlay of CBM42-4 in complex with the disaccharide (blue with green carbohydrate) and a Phyre2 [58] generated homology model of CBM32-3 (yellow) showing the conservation of residues lining the binding site. (G) and (H) Superposition of CBM32-4 in complex with the disaccharide (blue in both panels) with CBM32-5 in complex with galactose (yellow molecule in panel G) and with the GlcNAc binding CBM32 from *C. perfringens* NagH (PDB code 2W1U; grey molecule in panel H). Relevant residues involved in carbohydrate recognition are shown as sticks. Only residues in CBM32-4 are labeled.
doi:10.1371/journal.pone.0033524.g007

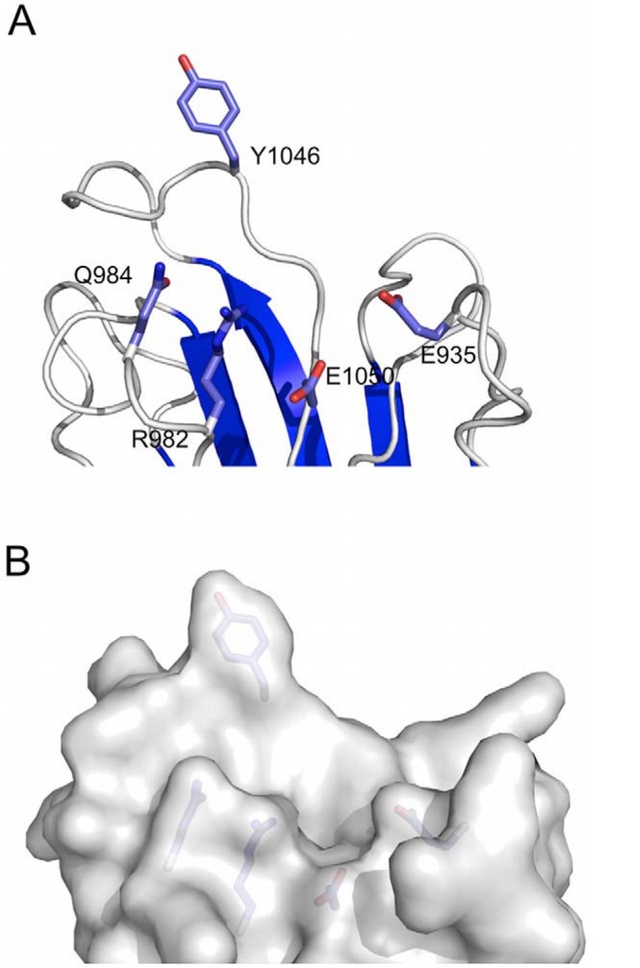


Figure 8. Homology model of CBM32-2 generated with Phyre2 [58]. (A) Architecture of the putative carbohydrate binding site with residues possibly involved in sugar recognition shown as sticks. (B) Solvent accessible surface of the putative binding site revealing its contours.
doi:10.1371/journal.pone.0033524.g008

galactose residues is quite low and on its own would be unlikely to mediate significant adherence of soluble CpGH89 to terminal galactose residues, the possible context of bacterial surface association of the entire enzyme creates further potential for avid binding.

Overall, the presence of at least three functional CBMs in CpGH89, with a fourth likely, imparts diversity in the ability of this enzyme to recognize carbohydrate substructures and potential for increased affinity through multivalent interactions. As a secreted enzyme this capability would enhance the overall association of the enzyme with class III mucins. In the possible case that CpGH89 is immobilized on the bacterial cell-surface the enzyme's capacity to bind carbohydrate would impart considerable carbohydrate-adhesive capacity to the bacterium thus promote the tight interaction of this bacterium with its host.

Materials and Methods

Cloning, protein production and purification

Gene fragments encoding desired CBMs from CpGH89 (locus tag CPF_0859) were PCR amplified from *C. perfringens* ATCC 13124 genomic DNA using oligonucleotide primers (see Table 2) with engineered 5′ and 3′ NheI and XhoI restriction endonuclease sites, respectively, incorporated into the ends of the primers. The following gene fragments were cloned into pET28a(+) through standard molecular biology procedures: CBM32-1 (nucleotides 76–462), CBM32-2 (nucleotides 2752–3171), CBM32-3 (nucleotides 3187–3603), CBM32-4 (nucleotides 3616–4029), CBM32-5 (nucleotides 4066–4479), CBM32-6 (nucleotides 4486–4863). All of the resulting gene fusions encoded an N-terminal six-histidine tag fused to the protein of interest by an intervening thrombin protease cleavage site. Bidirectional DNA sequencing was used to verify the fidelity of each construct.

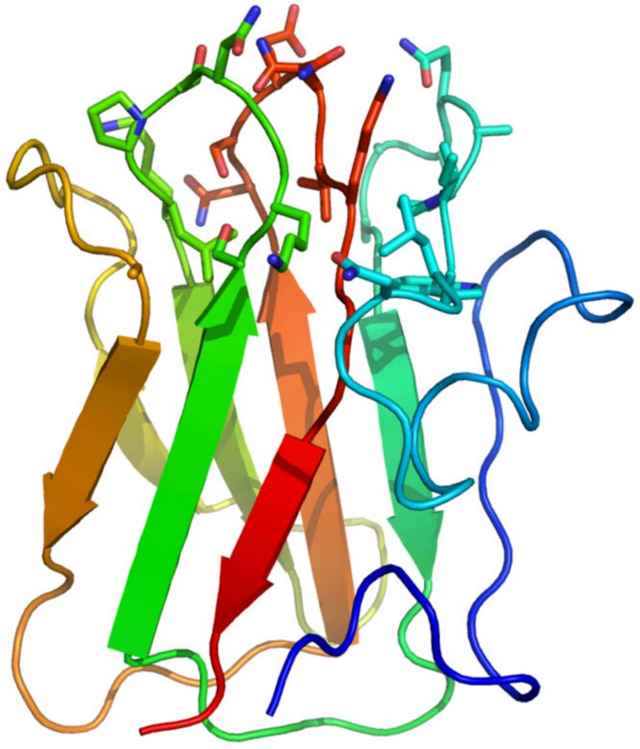


Figure 9. Structural analysis of CBM32-6. A cartoon representation of the structure of CBM32-6 determined by X-ray crystallography to 1.55 Å resolution. Amino acid side chains found in what is normally the carbohydrate-binding site of family 32 CBMs are shown in stick representation. This reveals the lack of side chains normally associated with carbohydrate binding, particularly a lack of aromatic amino acid side chains.
doi:10.1371/journal.pone.0033524.g009

All of the proteins were produced recombinantly in *E. coli* BL21(DE3) and purified by immobilized metal affinity chromatography and size exclusion chromatography (SEC) using methodologies described in detail previously [5]. Seleno-methionine-labeled CBM32-4 and CBM32-6 was produced as above using *E. coli* B834 (DE3) as the expression strain (Novagen). The media containing seleno-methionine was prepared according to the instructions of the manufacturer (Athena Enzyme). Protein concentrations were determined at 280 nm using calculated extinction coefficients [35] as follows: CBM32-1, 20340 M^{-1} cm^{-1}; CBM32-2, 17780 M^{-1} cm^{-1}; CBM32-3, 13940 M^{-1} cm^{-1}; CBM32-4, 15220 M^{-1} cm^{-1}; CBM32-5, 16500 M^{-1} cm^{-1}; CBM32-6, 15220 M^{-1} cm^{-1}.

Glycan microarray screening

Glycan microarray screening was performed by Core H of the Consortium for Functional Glycomics (www.functionalglycomics.

Table 2. Oligonucleotide primers used for amplification and cloning.

Oligonucleotide	Sequence	Used to amplify and clone
CBM32-1F	CAT ATG GCT AGC GGT GTT GAA ATT ACG GAA G	CBM32-1
CBM32-2F	CAT ATG GCT AGC GAA AGA GTT AAT ATT GCT	CBM32-2
CBM32-3F	CAT ATG GCT AGC GAA GAT GAG TAT ACT AAC G	CBM32-3
CBM32-4F	CAT ATG GCT AGC GCT AAT TAT GTA AAT ATA G	CBM32-4
CBM32-5F	CAT ATG GCT AGC GCA TTA CCT CAA GGA AAT	CBM32-5
CBM32-6F	CAT ATG GCT AGC GAA AAC CTA GCT ATG AAA G	CBM32-6
CBM32-1R	GAA TTC CTC GAG TTA ACC AAA TAC ATT TAT TTC	CBM32-1
CBM32-2R	GAA TTC CTC GAG TTA ATA TAC CAT TAT TTC TGC	CBM32-2
CBM32-3R	GAA TTC CTC GAG TTA TGA CAT GGC CTT TAC TTC	CBM32-3
CBM32-4R	GAA TTC CTC GAG TTA ACT CAT AGC TTT AAT TTC	CBM32-4
CBM32-5R	GAA TTC CTC GAG TTA TGC AAA TAC ATT TAA TTC	CBM32-5
CBM32-6R	GAA TTC CTC GAG TTA TCC TTT ATA AAT TTT GAT	CBM32-6

doi:10.1371/journal.pone.0033524.t002

org/). CBMs were labeled by coupling to Alexa Fluor® 488 labeled streptavidin *via* a biotin-NTA:Ni^{2+} linker using methods identical to those described previously [36]. Labeled proteins were desalted using PD-10 columns (GE Healthcare) and used to probe the printed glycan arrays according to the standard procedures of Core H of the Consortium for Functional Glycomics.

Binding studies

Qualitative UV difference scans were performed using methods identical to those described previously [31]. Quantitative UV difference titrations were also performed using methods already described [27]. The concentration of protein used for the titrations was 31.5 μM in 20 mM Tris-HCl pH 8.0. The concentrations of carbohydrate stocks used to titrate into protein varied between ~40 mM and 45 mM and were prepared by mass in 20 mM Tris-HCl pH 8.0. Experiments were performed at 25°C in triplicate.

Isothermal Titration Calorimetry was performed as described previously using a VP-ITC (MicroCal, Northampton, MA)[27]. Proteins were filtered and degassed prior to use. Carbohydrate solutions were prepared by mass in buffer saved from dialysis of the appropriate protein. These solutions were also filtered and degassed prior to use. The proteins concentrations used varied from ~100 μM to ~550 μM. However, in no case could a protein concentration be used that exceeded the K_d by more than five-fold (i.e. C-values were less than 5), thus, data was fit with a single binding site model using MicroCal Origin software (version 7.0) with the stoichiometry (*n*-value) fixed at 1. Experiments using CBM32-5 were performed in 20 mM Tris-HCl, pH 8.0, and those with CBM32-3 and CBM32-4 in 50 mM HEPES, pH 7.5. Experiments were performed at 25°C in triplicate.

Crystallization

Prior to crystallization, CBMs generally required overnight treatment with thrombin followed by re-purification by SEC to remove the 6-histidine tag. The complex of CBM32-4 with GlcNAc-α-1,4-Gal, however, was obtained with protein still having the 6-histidine tag. All crystallization experiments were performed at 18°C using the hanging drop vapour diffusion method.

Seleno-methionine labeled CBM32-4 at 15 mg/ml crystallized in 0.2 M KSCN, 22% polyethylene glycol (PEG) 3350, 0.1 M Tris-HCL pH 7.5. 20% ethylene glycol in crystallization solution was used as a cryoprotectant. Unlabeled CBM32-4 (20 mg/ml) in complex with GlcNAc-α-1,4-Gal (at 2 mM) crystallized in 0.1 M ZnOAc, 0.1 M Bicine pH 8.0, 18% PEG 3350, 4 mM CrCl; 20% ethylene glycol in this crystallization solution was used as a cryoprotectant.

All crystals of CBM32-5 were obtained using the protein at 20 mg/ml. Complexes were obtained by co-crystallization of the protein with the carbohydrate under the following conditions: the galactose (10 mM) and GalNAc (10 mM) complexes crystallized in 0.1 M Bis-Tris pH 5.5, 20% PEG 4000, and 0.2 M $LiSO_4$; the GalNAc-β-1-3Gal (10 mM) complex crystallized in 0.1 M NaCitrate pH 5.6, 20% PEG 3350, and 0.2 M MgOAc; the Tn Antigen [10 mM; *N*-acetyl-α-D-galactosaminyl-1-O-serine (V-labs)] complex crystallized in 0.1 M NaCitrate pH 5.6, 20% PEG 3350, and 0.1 M ZnOAcetate. In all cases the crystals were cryoprotected using the crystallization solution supplemented with 15% glycerol.

Seleno-methionine labeled CBM32-6 (20 mg/ml) was crystallized in 0.1 M Bis-Tris pH 6.5, 29% PEG 3350, 0.05 M CaCl2 and 20% ethylene glycol in crystallization solution was used for cryoprotection.

Data collection, Structure Solution and Refinement

Diffraction data were collected at 100 K at the National Synchrotron Light Source (NSLS) beamline X8-C, the Stanford Synchrotron Radiation Laboratories (SSRL) beamline BL 9-2, or a home source comprising a Rigaku R-AXIS IV++ area detector coupled to a MM-002 X-ray generator with Osmic "blue" optics and Oxford Cryostream 700 as indicated in Tables 3 and 4. Data were processed using d*trek or MOSFLM [37,38].

The structures of CBM32-4 and CBM32-6 were solved by single-anomalous dispersion (SAD) experiments optimized for selenium (see Table 4 for wavelengths at which SAD data were collected). The heavy atom substructures were determined from the SAD data using the program ShelXC/D, while phasing was performed using ShelxE [39]. CBM32-4 crystallized with a single molecule in the AU; three of its potential four selenium sites were found and used for phasing. CBM32-6 crystallized with a two molecules in the AU with each monomer having two potential selenium sites; only one selenium site per monomer was found and used for phasing. Density modification with the program DM [40,41] was used to improve the phases prior to model building.

Table 3. X-ray data collection and model refinement statistics for CBM32-5.

Data collection statistics	**CBM32-5 galactose**	**CBM32-5 galNAc**	**CBM32-5 TnAg**	**CBM32-5 galNac-β-1,3-gal**
Wavelength	1.5418	1.5418	1.5418	1.5418
Beamline	MM-002	MM-002	MM-002	MM-002
Space group	C2	C2	$P2_12_12_1$	$P2_12_12_1$
Resolution	20.00-1.55 (1.59-1.55)	30.00-1.90 (1.95-1.90)	20.00-1.70 (1.74-1.70)	20.00-1.75 (1.80-1.75)
Cell dimension α, β, γ (Å)	65.27, 38.32, 53.88 90.00 90.64 90.00	65.82, 37.27, 57.41 90.00,103.51,90.00	31.80, 59.26, 67.43 90.00, 90.00, 90.00	33.80, 56.31, 70.66 90.00, 90.00, 90.00
R_{merge}	0.061 (0.0244)	0.067 (0.323)	0.059 (0.314)	0.064 (0.378)
Completeness (%)	99.6 (99.5)	99.3 (97.8)	97.5 (95.2)	97.9 (98.0)
$<I/\sigma I>$	13.0 (2.1)	11.6 (3.9)	12.0 (3.5)	10.7 (3.2)
Redundancy	3.1 (2.5)	6.1 (6.0)	4.3 (3.8)	4.4 (4.1)
Total reflections	66379	66214	61394	63956
Unique reflections	21421	10778	14287	14448
Refinement statistics				
R (%)	18.6	20.4	18.8	19.9
R_{free} (%)	22.4	26.5	22.2	24.6
RMSD				
Bond lengths (Å)	0.013	0.014	0.012	0.015
Bond angles (°)	1.415	1.380	1.801	1.716
Average *B*-factors (Å^2)				
Protein Chain	13.6	28.3	16.4	24.5
Water molecules	28.7	35.5	31.4	33.9
Ligand molecules	15.3	29.0	17.6	58.5
Number of atoms				
Protein atoms Chain A	1081	1076	1077	1063
Water molecules	273	141	227	155
Ligand molecules	12	15	21	26
Ramachandran statistics				
Most favored (%)	97.9	96.5	96.4	99.3
Additional allowed (%)	1.4	3.5	3.6	0.7
Disallowed (%)	0.7	0	0	0

doi:10.1371/journal.pone.0033524.t003

ARP/wARP [42] was able to build almost complete models, which were completed by manual model building with COOT [43]. Structural refinement of CBM32-6 (selenium derivative) was performed with PHENIX [44] refine using simulated annealing interspersed with manual building in COOT [43]. REFMAC [45] was used to refine CBM32-4. The structure of CBM32-4 in complex with GlcNAc-α-1,4-Gal was solved by molecular replacement using PHASER [46] to find the two molecules in the asymmetric unit. The model was completed by manual building with COOT and refinement with REFMAC; TLS parameters were included in the final refinement cycles of this structure.

The structure of CBM32-5 in complex with galactose was solved by molecular replacement using CpCBM32C from CpGH84C as a search model (PDB id 2j1e [27]) and MOLREP [47] to find the single molecule in the asymmetric unit. Automated model building was carried out with ARP/wARP followed by manual completion with COOT. This structure was used as a starting point to solve the structures of CBM32-5 in complex with other sugars. All refinements were carried out using REFMAC.

In all cases, waters were added using COOT:FINDWATERS. In all datasets 5% of the observations were flagged as "free" and used to monitor refinement progress. Final models were validated with MOLPROBITY [48]. Tables 3 and 4 show the data collection, refinement and final model validation statistics.

Modeling the CBM32-5 tetrasaccharide complex

A 50 ns molecular dynamics (MD) simulation of the tetrasaccharide, GalNAcα1-4(Fucα1-2)Galβ1-4GlcNAcβ with a reducing terminal methyl, was performed using the pmemd module of the AMBER11 software package [49]. The GLYCAM06g [50] force field was used for the tetrasaccharide parameters while the initial geometry was obtained from the GLYCAM carbohydrate 3D structure web tool [51]. The tetrasaccharide was explicitly solvated with 1724 TIP3P waters [52] and no ions. Minimization was performed for 20,000 steps, half of which used the conjugate gradient method followed by the steepest descent method. A 10.050 ns constant pressure MD (NPT) was used to ensure water and glycan equilibration in which the first 50 ps were used to heat the system from 5 K to 300 K. The final frame from equilibration

Table 4. X-ray data collection and model refinement statistics for CBM32-4 and CBM32-6.

Data collection statistics	**CBM32-4** **Seleno-Methionine**	**CBM32-4** **glcNAc-α-1,4-gal**	**CBM32-6** **Seleno-Methionine**
Wavelength	0.9796	1.5418	0.9790
Beamline	NSLS X8C	MM-002	SSRL BL9-2
Space group	$P4_{3}2_{1}2$	$P2_{1}2_{1}2$	$P4_{3}$
Resolution	20.00-1.55 (1.64-1.55)	20.00-2.80 (2.87-2.80)	30.00-1.55 (1.63-1.55)
Cell dimension α, β, γ (Å)	53.20, 53.20, 110.60 90.0, 90.0, 90.0	90.0, 90.0, 90.0 89.71, 49.89, 63.17	90.0, 90.0, 90.0 48.81, 48.81, 98.18
R_{merge}	0.107 (0.403)	0.143 (0.329)	0.048 (0.377)
Completeness (%)	99.9 (99.8)	93.4 (90.5)	100.0 (100.0)
$\langle I/\sigma I \rangle$	18.9 (7.6)	5.6 (2.5)	30.6 (7.6)
Redundancy	16.3 (16.6)	3.6 (3.8)	15.2 (15.2)
Total reflections	389006	25312	505862
Unique reflections	23846	6965	33325
Refinement statistics			
R (%)	12.8	28.6	19.6
R_{free} (%)	17.3	31.8	23.9
RMSD			
Bond lengths (Å)	0.018	0.006	0.011
Bond angles (°)	1.649	1.053	1.294
Average *B*-factors (Å^2)			
Protein Chain A	11.1	24.9	22.7
Protein Chain B	N/A	19.6	29.6
Water molecules	30.7	23.7	32.5
Ligand	N/A	69.5 (A); 33.1 (B)	N/A
Number of atoms			
Protein atoms Chain A	1064	1097	1017
Protein atoms Chain B	N/A	1103	982
Water molecules	261	111	174
Ligand	N/A	52	N/A
Ramachandran statistics			
Most favored (%)	95.7	92.6	96.9
Additional allowed (%)	4.3	6.3	1.3
Disallowed (%)	0	1.1	1.8

doi:10.1371/journal.pone.0033524.t004

was used to start the 50 ns NPT production simulation of the tetrasaccharide. In all tetrasaccharide simulations an 8.0 Å van der Waals cutoff was employed, particle mesh Ewald summation (PME)[53] was used for long range electrostatics, 1,4-scaling factors were set to unity, and a dielectric of 1.0 were employed. The Berendsen thermostat was used with a coupling constant of 1.0 ps. Pressure was maintained at 1 atm with a relaxation time of 0.1 ps. The SHAKE [54] algorithm was used to restrain the bonds to hydrogens reducing the time between steps to 2 fs. Production frames were collected at every ps and only the production run was used for further analyses.

The crystal structure of GalNAc-β-Serine bound to the CBM32-5 was used as a template for modeling the tetrasaccharide onto the complex. The GalNAc-β from the template crystal structure and the non-reducing terminal GalNAc-β from the MD simulation were aligned on the ring atoms (C1, C2, C3, C4, C5 and O5) using the alignment algorithm in VMD [55]. Then the MD trajectory of the entire tetrasaccharide was combined together with the template protein coordinates resulting in 50,000 snapshots of the solution tetrasaccharide bound to the crystal protein coordinates. Clashes were removed using a 2,000 step minimization, half conjugate gradient and half steepest descent, for each of the 50,000 complexes where the FF99SB force field [56] was used for the protein. The modified Onufriev, Bashford and Case generalized Borne implicit solvent was used [57] to approximate solvent effects in minimization. All minimizations in developing the CBM-tetrasaccharide complexes used mixed 1,4-scaling, which set van der Waals and electrostatic scaling factors to 1.2 and 2.0, respectively, for the protein (consistent with FF99SB) and unity for the tetrasaccharide (consistent with GLYCAM06). Additionally, a 12.0 Å long-range van der Waals cutoff was employed with PME being used for long-range electrostatics.

The final net energy (including GB solvation contributions) of the CBM-tetrasaccharide complex was used to identify complexes within 15 kcal/mol of the lowest energy complex. This resulted in the selection of 42 complexes, which were further minimized using

10,000 steps of conjugate gradient and 10,000 steps of steepest descent minimization. These new complexes were then ranked according to their overall system energy and grouped together using a 1.0 Å cutoff in root mean squared deviation of the heavy atoms. The models were grouped such that reference structures were selected starting from the lowest energy and ending at the highest energy models. Structures grouped from the lowest energy clusters were excluded from subsequent root mean square deviation grouping analyses meaning any single representation could only belong to one group. Ten clusters were identified in which 60% of the complexes were in the two lowest energy groupings, 33% in the lowest energy group. Energy decomposition was performed on these ten clusters using the MMGBSA.py application in AMBER using the same implicit solvent model as in the minimizations.

Homology modeling of CBM32-3 and CBM32-2

Structural models of CBM32-3 and CBM32-2 were prepared using the one-to-one threading function of the Phyre2 server [58]. In both cases, the 1.55 Å resolution structure of CBM32-4 was used as a template.

Accession Codes

Coordinates and structure factors have been deposited in the protein data bank with the following accession codes: **4a3z** for CBM32-4 (seleno-methionine labeled), **4a6o** for CBM32-4 in complex with GlcNAc-α-1,4-Gal, **4a41** for CBM32-5 in complex with galactose, **4aax** for CBM32-5 in complex with GalNAc, **4a45** for CBM32-5 in complex with GalNAc-β-1,3-Gal, **4a44** for CBM32-5 in complex with the Tn Antigen, and **4a42** for CBM32-6 (seleno-methionine labeled).

Supporting Information

Figure S1 The energy decomposition profiles of residues within 5.0 Å of the tetrasaccharide, GalNAc-α-1,4(Fuc-α-1,2)-Gal-β-1,4-GlcNAc, modeled onto the crystal structure of CBM32-5. The non-polar contributions (top), polar contributions (middle), and net binding contributions (bottom) are shown on a per-residue basis. While the predominant interaction is between the protein and GalNAc, the fucose adds significant non-polar contributions to the binding through residues Y1395 and N1396.
(DOC)

Acknowledgments

We are grateful to Core H of the Consortium for Functional Glycomics for performing the glycan array experiments. We thank the staff of the SSRL beamline 9-2 and the NSLS beamline X8C. ABB is a Canada Research Chair in Molecular Interactions and a Michael Smith Foundation for Health Research Scholar.

Author Contributions

Conceived and designed the experiments: EFB ABB MT RJW. Performed the experiments: EFB CPS MDS MC MT RJW ABB. Analyzed the data: EFB MDS MT RJW ABB. Wrote the paper: EFB ABB.

References

1. McGuckin MA, Linden SK, Sutton P, Florin TH (2011) Mucin dynamics and enteric pathogens. Nat Rev Microbiol 9: 265–278.
2. Atuma C, Strugala V, Allen A, Holm L (2001) The adherent gastrointestinal mucus gel layer: thickness and physical state in vivo. Am J Physiol Gastrointest Liver Physiol 280: G922–929.
3. Boraston AB, Ficko-Blean E, Healey M (2007) Carbohydrate recognition by a large sialidase toxin from *Clostridium perfringens*. Biochemistry 46: 11352–11360.
4. Newstead SL, Potter JA, Wilson JC, Xu G, Chien CH, et al. (2008) The structure of *Clostridium perfringens* NanI sialidase and its catalytic intermediates. J Biol Chem 283: 9080–9088.
5. Ficko-Blean E, Boraston AB (2005) Cloning, recombinant production, crystallization and preliminary X-ray diffraction studies of a family 84 glycoside hydrolase from *Clostridium perfringens*. Acta Crystallogr Sect F Struct Biol Cryst Commun 61: 834–836.
6. Ficko-Blean E, Gregg KJ, Adams JJ, Hehemann JH, Czjzek M, et al. (2009) Portrait of an enzyme, a complete structural analysis of a multimodular {beta}-N-acetylglucosaminidase from *Clostridium perfringens*. J Biol Chem 284: 9876–9884.
7. Abbott DW, Eirin-Lopez JM, Boraston AB (2008) Insight into ligand diversity and novel biological roles for family 32 carbohydrate-binding modules. Mol Biol Evol 25: 155–167.
8. Koutsioulis D, Landry D, Guthrie EP (2008) Novel endo-alpha-N-acetylgalactosaminidases with broader substrate specificity. Glycobiology 18: 799–805.
9. Ashida H, Maki R, Ozawa H, Tani Y, Kiyohara M, et al. (2008) Characterization of two different endo-alpha-N-acetylgalactosaminidases from probiotic and pathogenic enterobacteria, *Bifidobacterium longum* and *Clostridium perfringens*. Glycobiology 18: 727–734.
10. Ficko-Blean E, Stubbs KA, Nemirovsky O, Vocadlo DJ, Boraston AB (2008) Structural and mechanistic insight into the basis of mucopolysaccharidosis IIIB. Proc Natl Acad Sci U S A 105: 6560–6565.
11. Fujita M, Tsuchida A, Hirata A, Kobayashi N, Goto K, et al. (2011) Glycoside hydrolase family 89 alpha-N-acetylglucosaminidase from *Clostridium perfringens* specifically acts on GlcNAc alpha1,4Gal beta1R at the non-reducing terminus of O-glycans in gastric mucin. J Biol Chem 286: 6479–6489.
12. Kochetkov NK, Derevitskaya VA, Arbatsky NP (1976) The structure of pentasaccharides and hexasaccharides from blood group substance H. Eur J Biochem 67: 129–136.
13. Ota H, Katsuyama T (1992) Alternating laminated array of two types of mucin in the human gastric surface mucous layer. Histochem J 24: 86–92.
14. Akamatsu T, Katsuyama T (1990) Histochemical demonstration of mucins in the intramucosal laminated structure of human gastric signet ring cell carcinoma and its relation to submucosal invasion. Histochem J 22: 416–425.
15. Tsutsumi Y, Nagura H, Osamura Y, Watanabe K, Yanaihara N (1984) Histochemical studies of metaplastic lesions in the human gallbladder. Arch Pathol Lab Med 108: 917–921.
16. Nakajima K, Ota H, Zhang MX, Sano K, Honda T, et al. (2003) Expression of gastric gland mucous cell-type mucin in normal and neoplastic human tissues. Journal of Histochemistry & Cytochemistry 51: 1689–1698.
17. Matsuzawa K, Akamatsu T, Katsuyama T (1992) Mucin Histochemistry of Pancreatic Duct Cell-Carcinoma, with Special Reference to Organoid Differentiation Simulating Gastric Pyloric Mucosa. Human Pathology 23: 925–933.
18. Ota H, Hayama M, Nakayama J, Hidaka H, Honda T, et al. (2001) Cell lineage specificity of newly raised monoclonal antibodies against gastric mucins in normal, metaplastic, and neoplastic human tissues and their application to pathology diagnosis. Am J Clin Pathol 115: 69–79.
19. Zhang MX, Nakayama J, Hidaka E, Kubota S, Yan J, et al. (2001) Immunohistochemical demonstration od alpha,4-N-acetylglucosaminyltransferase that forms GlcNAc alpha 1,4Gal beta residues in human gastrointestinal mucosa. Journal of Histochemistry & Cytochemistry 49: 587–596.
20. Lesuffleur T, Zweibaum A, Real FX (1994) Mucins in normal and neoplastic human gastrointestinal tissues. Crit Rev Oncol Hematol 17: 153–180.
21. Fujimori Y, Akamatsu T, Ota H, Katsuyama T (1995) Proliferative markers in gastric carcinoma and organoid differentiation. Human Pathology 26: 725–734.
22. Nakamura N, Ota H, Katsuyama T, Akamatsu T, Ishihara K, et al. (1998) Histochemical reactivity of normal, metaplastic, and neoplastic tissues to alpha-linked N-acetylglucosamine residue-specific monoclonal antibody HIK1083. J Histochem Cytochem 46: 793–801.
23. Kijima H, Watanabe H, Iwafuchi M, Ishihara N (1989) Histogenesis of gallbladder carcinoma from investigation of early carcinoma and microcarcinoma. Acta Pathol Jpn 39: 235–244.
24. Nakayama J, Yeh JC, Misra AK, Ito S, Katsuyama T, et al. (1999) Expression cloning of a human alpha1, 4-N-acetylglucosaminyltransferase that forms GlcNAcalpha1→4Galbeta→R, a glycan specifically expressed in the gastric gland mucous cell-type mucin. Proc Natl Acad Sci U S A 96: 8991–8996.
25. Kawakubo M, Ito Y, Okimura Y, Kobayashi M, Sakura K, et al. (2004) Natural antibiotic function of a human gastric mucin against *Helicobacter pylori* infection. Science 305: 1003–1006.
26. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J 382: 769–781.
27. Ficko-Blean E, Boraston AB (2006) The interaction of a carbohydrate-binding module from a *Clostridium perfringens* N-acetyl-beta-hexosaminidase with its carbohydrate receptor. J Biol Chem 281: 37748–37757.
28. Gaskell A, Crennell S, Taylor G (1995) The three domains of a bacterial sialidase: a beta-propeller, an immunoglobulin module and a galactose-binding jelly-roll. Structure 3: 1197–1205.

29. Boraston AB, Warren RA, Kilburn DG (2001) beta-1,3-Glucan binding by a thermostable carbohydrate-binding module from *Thermotoga maritima*. Biochemistry 40: 14679–14685.
30. Boraston AB, Notenboom V, Warren RA, Kilburn DG, Rose DR, et al. (2003) Structure and ligand binding of carbohydrate-binding module CsCBM6-3 reveals similarities with fucose-specific lectins and ''galactose-binding'' domains. J Mol Biol 327: 659–669.
31. Ficko-Blean E, Boraston AB (2009) N-acetylglucosamine recognition by a family 32 carbohydrate-binding module from *Clostridium perfringens* NagH. J Mol Biol 390: 208–220.
32. Newstead SL, Watson JN, Bennet AJ, Taylor G (2005) Galactose recognition by the carbohydrate-binding module of a bacterial sialidase. Acta Crystallogr D Biol Crystallogr 61: 1483–1491.
33. Johansson ME, Larsson JM, Hansson GC (2011) The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. Proc Natl Acad Sci U S A 108 Suppl 1: 4659–4665.
34. Johansson ME, Phillipson M, Petersson J, Velcich A, Holm L, et al. (2008) The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. Proc Natl Acad Sci U S A 105: 15064–15069.
35. Gasteiger E, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. (2005) Protein Identification and Analysis Tools on the ExPASy Server. In: Walker JM, ed. The Proteomics Protocols Handbook: Humana Press. pp 571–607.
36. Higgins MA, Ficko-Blean E, Meloncelli PJ, Lowary TL, Boraston AB (2011) The overall architecture and receptor binding of pneumococcal carbohydrate-antigen-hydrolyzing enzymes. J Mol Biol 411: 1017–1036.
37. McCalmont TH (2011) Crystal clear. J Cutan Pathol 38: 540–541.
38. Powell HR (1999) The Rossmann Fourier autoindexing algorithm in MOSFLM. Acta Crystallogr D Biol Crystallogr 55: 1690–1695.
39. Schneider TR, Sheldrick GM (2002) Substructure solution with SHELXD. Acta Crystallographica Section D-Biological Crystallography 58: 1772–1779.
40. Cowtan K (1994) DM: An automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography. pp 34–38.
41. Cowtan K, Main P (1998) Miscellaneous algorithms for density modification. Acta Crystallogr D Biol Crystallogr 54: 487–493.
42. Morris RJ, Perrakis A, Lamzin VS (2002) ARP/wARP's model-building algorithms. I. The main chain. Acta Crystallogr D Biol Crystallogr 58: 968–975.
43. Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 60: 2126–2132.
44. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr 66: 213–221.
45. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53: 240–255.
46. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, et al. (2007) Phaser crystallographic software. Journal of Applied Crystallography 40: 658–674.
47. Vagin A, Teplyakov A (1997) MOLREP: an automated program for molecular replacement. Journal of Applied Crystallography 30: 1022–1025.
48. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, et al. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66: 12–21.
49. Case DA, Darden TA, Cheatham III TE, Simmerling CL, Wang J, et al. (2010) AMBER11. University of California, San Francisco, CA.
50. Kirschner KN, Yongye AB, Tschampel SM, Gonzalez-Outeirino J, Daniels CR, et al. (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates. J Comput Chem 29: 622–655.
51. Woods Group (2005–2012) Complex Carbohydrate Research Center, The University of Georgia, Athens, GA. (http://www.glycam.com).
52. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. Journal of Chemical Physics 79: 926–935.
53. Darden T, York D, Pedersen L (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. Journal of Chemical Physics 98: 10089–10092.
54. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. Journal of Computational Physics 23: 327–341.
55. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. Journal of Molecular Graphics 14: 33–38.
56. Simmerling C, Hornak V, Abel R, Okur A, Strockbine B, et al. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins-Structure Function and Bioinformatics 65: 712–725.
57. Case DA, Onufriev A, Bashford D (2000) Modification of the generalized Born model suitable for macromolecules. Journal of Physical Chemistry B 104: 3712–3720.
58. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nature Protocols 4: 363–371.
59. Read RJ (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Crystallographica Section A A42: 140–149.