JBC Papers in Press. Published on July 22, 2015 as Manuscript M115.660878
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.M115.660878

Structural aspects of N-glycosylations and the C-terminal region in human glypican-1

Wael Awad[a,b], Barbara Adamczyk[c], Jessica Örnros[c], Niclas G. Karlsson[c], Katrin Mani[d], Derek T. Logan[a]*

[a]Department of Biochemistry and Structural Biology, Centre for Molecular Protein Science, Lund University, Getingevägen 60, Box 124, SE-221 00 Lund, Sweden.
[b]Department of Biophysics, Faculty of Science, Cairo University, Cairo, Egypt.
[c]Medical Biochemistry, University of Gothenburg, Gothenburg, Sweden
[d]Department of Experimental Medical Science, Division of Neuroscience, Glycobiology Group, Lund University, Lund, Sweden.

*Corresponding author: phone, +4646-2221443; fax: +4646-2224116; e-mail: derek.logan@biochemistry.lu.se

Running title: Solution structure of human glypican-1 core protein

Keywords: glypicans; glypican-1; membrane-anchored protein; proteoglycan; small-angle X-ray scattering (SAXS); mass spectrometry (MS); N-linked glycosylation; structure-function

Background: Glypicans are a family of cell surface proteoglycans implicated in diverse cell signalling pathways.
Results: Description of the structure and functions of the N-glycans and C-terminus of human glypican-1.
Conclusions: The studies revealed the structural topology of glypicans with respect to the membrane and the protective roles of their N-glycans.
Significance: Improved structural knowledge on glypican-1 helps to elucidate the functional roles of the glypicans.

SUMMARY

Glypicans are multifunctional cell surface proteoglycans involved in several important cellular signalling pathways. Glypican-1 (Gpc1) is the predominant heparan sulphate (HS) proteoglycan in the developing and adult human brain. The two N-linked glycans and the C-terminal domain that attaches the core protein to the cell membrane are not resolved in the Gpc1 crystal structure. Therefore we have studied Gpc1 using crystallography, small-angle X-ray scattering and chromatographic approaches to elucidate the composition, structure and function of the N-glycans and the C-terminus, and also the topology of Gpc1 with respect to the membrane. The C-terminus is shown to be highly flexible in solution, but it orients the core protein transverse to the membrane, directing a surface evolutionarily conserved in Gpc1 orthologues towards the membrane, where it may interact with signalling molecules and/or membrane receptors on the cell surface, or even the enzymes involved in HS substitution in the Golgi apparatus Furthermore, the N-glycans are shown to extend the protein stability and lifetime by protection against proteolysis and aggregation.

INTRODUCTION

Glypiated proteins are anchored to the extracellular surface of the eukaryotic cell membrane by covalent linkage of their C-termini to glycosyl-phosphatidylinositol (GPI) (1). Glypicans (Gpcs) are a family of glypiated extracellular proteoglycans that mainly work as coregulators of several signalling pathways, and are thereby involved in the control of many biological processes such as cellular division, differentiation, and morphogenesis. To date, six different Gpcs have been identified in vertebrates, Gpc1 to Gpc6, with approximately 25% amino acid identity within the family, two in *C. elegans* (Gpc1 and Lon-2), two in *Drosophila melanogaster* (Dally and Dally-like

protein) and one in zebrafish (knypek). The mature forms of human Gpcs have a core protein of ~60–70 kDa in size and share a pattern of 14 conserved cysteine residues. In the C-terminal region, Gpcs share attachment sites for glycosaminoglycan (GAG) chains and a hydrophobic sequence for the addition of the GPI anchor (2).

GAGs are O-linked linear anionic hetero-polysaccharides found in all mammalian tissues in extracellular and/or intracellular environments. Their sequences are composed of repeating disaccharide units and are divided into classes depending on the disaccharide building blocks. The most widespread categories are: 1) heparan sulphate (HS), which consists of repeating units of glucuronic acid (GlcA) or iduronic acid (IdoA) followed by N-acetylglucosamine (GlcNAc), i.e. $(GlcA/IdoA\text{-}GlcNAc)_n$; 2) chondroitin sulphate (CS) and dermatan sulphate (DS), which consist of GlcA or IdoA followed by N-acetylgalactosamine (GalNAc), i.e. $(GlcA/IdoA\text{-}GalNAc)_n$. The biosynthesis of GAG chains is performed by membrane-bound glycosyltransferases in the Golgi apparatus and starts with addition of xylose (Xyl) to the GAG attachment serine residue by a xylosyltransferase. Then two consecutive galactose units and a GlcA unit are transferred, forming the linkage tetrasaccharide (GlcAβ1-3Galβ1-3Galβ1-4Xylβ1-O-Ser).

Addition of the next residue is the critical factor in determining which type of GAG will be formed: incorporation of GalNAc promotes CS/DS assembly and addition of GlcNAc initiates HS synthesis. The GAG chain then becomes elongated by addition of the corresponding repeating disaccharides and undergoes serial modifications including N-deacetylation, N- & O-sulphation and epimerization, generating GAG chains with heterogeneous structures (3).

Human glypican-1 (Gpc1) is mainly expressed in the neural and skeletal systems during development, as well as many other tissues in the adult. Gpc1 is involved in the uptake of different macromolecules such as growth factors (FGF2), viral proteins, cytokines and polyamines (4). Gpc1 knockout mice have significant reduction in brain size at birth (about 30%), indicating a role for Gpc1 in brain development (5). Many studies have revealed involvement of Gpc1 in the pathogenesis of neurodegenerative diseases, including Alzheimer's disease and transmissible spongiform encephalopathies (6,7). Moreover, it has been shown that Gpc1 is upregulated in many human cancer types such as glioma, pancreatic cancer and breast cancer (8,9).

The Gpc1 protein is composed of an N-terminal core (residues: 24-474) and a C-terminal GAG attachment region (residues: 475-530), ending with a sequence of hydrophobic residues that link it to the GPI anchor (Figure 1A). The Gpc1 protein is decorated with two N-linked glycans at positions N79 and N116 and is also substituted with three chains of HS at S486, S488 and S490 (Figure 1A). Previously, we determined the crystal structure of the N-glycosylated, C-terminally truncated Gpc1 core protein at 2.6 Å (PDB entry 4ACR) and showed that it has an elongated, cylindrical form with dimensions 120 x 30 x 30 Å and an all α-helical fold (α1- α14) with three major loops (L1, L2 and L3) (Figure 1B) (10). The complete disulphide pattern of the 14 conserved Cys residues across the Gpc family was also revealed in the structure. Six disulphide bonds are located near the protein N-terminus in a region termed the Cys-rich lobe. This region is followed by the central lobe, stabilized by two evolutionary conserved hydrophobic cores. Finally, the protease site lobe contains a furin protease site found in many Gpc family members. Recently, we achieved improvements of Gpc1 crystal diffraction properties by controlled crystal dehydration using a humidity control device (HC1b) (11) and generated better electron density for crystals of C-terminally truncated Gpc1 (12). Unfortunately, the 3 Å crystal structure of full-length Gpc1 (PDB entry 4AD7) contains no electron density for the C-terminal region (approximately 53 residues) that attaches the core protein to the cell membrane. Therefore it was not possible to predict the location and orientation of the Gpc1 core protein and its HS chains with respect to the cell surface. Previous work from our laboratory shows that Gpc1 is invariably decorated with two N-glycans, which affect the Gpc1 expression level and HS substitution (13). Interestingly, the Gpc1 protein was correctly folded in the absence of N-glycans. The crystal structure of Gpc1 displayed only the first residue of the N-linked glycans and therefore it is still of great interest to characterize the structure and function of Gpc1 N-glycans using other techniques.

In this work, we couple the existing knowledge from protein crystallography with small angle X-ray scattering (SAXS) and other biophysical

techniques to provide new insights into the structural-functional characteristics of the Gpc1 core protein in solution. Briefly, we report the most high resolution and complete full-length Gpc1 crystal structure to date, by further application of crystal dehydration to this system. Moreover, the composition and structure of the N-linked glycans has been characterized. Finally, molecular modelling based on small-angle X-ray scattering (SAXS) data from various Gpc1 constructs was pursued to elucidate the spatial relationship between the cell membrane, Gpc1 core protein, and the N- and O-linked glycans. Taken together our study provides a clearer topological picture of Gpc1 with respect to the cell membrane.

EXPERIMENTAL PROCEDURES

Expression and purification of human Gpc1: The design of different Gpc1 variants is summarized in Table 1. All Gpc1 constructs were cloned and expressed in stable HEK293 cells as described previously (13). The conditioned medium was collected and the proteins were purified using Ni-NTA affinity chromatography. The proteins were then dialyzed into 20 mM NaCl, 20 mM Tris (pH 8) for the C-terminally truncated protein (Gpc1-dC) and the same buffer with additional 2mM DTT for the others before any further experiments.

Enzymatic treatments: Deglycosylation of Gpc1-dC and the full-length protein without the HS chains (Gpc1-dHS) was carried out by growing the Gpc1-producing cells in protein-free medium containing 10 mM of the plant alkaloid kifunensine, which forces the cells to express proteins with N-glycans of the high-mannose type that are sensitive to endoglycosidase H enzyme (EndoH) (14). Afterwards, EndoH treatment (New England Biolabs) of 1 mg of native high mannose Gpc1 was carried out in 25mM sodium phosphate buffer pH 7.0 through overnight incubation at 37 °C with 30 mU of the enzyme. The deglycosylation efficiency was tested by SDS-PAGE, which showed a reduction in the size of the Gpc1-dC & Gpc1-dHS core protein bands by 5 kDa after deglycosylation, producing deglycosylated forms of the Gpc1 core proteins (Gpc1-dC-dN and Gpc1-dHS-dN) respectively. EndoH was removed from the sample by repeating the Ni-NTA purification.

Digestion of the HS chains was carried out as follows: the purified proteins were dialyzed into 10 mM HEPES and 3mM $Ca(OAc)_2$ buffer pH 7.0 overnight, then 150 mU of HS lyase (Seikagaku, Japan) was added to 1 mg protein and incubated overnight at 37 °C. The protein buffer was exchanged to 0.3 M NaCl, 50 mM Na phosphate pH 8.0 by ultrafiltration, then incubated with DE-53 DEAE cellulose (for anion exchange) on a rocker at 8 °C for 1 h. The unbound proteins (without heparan sulphate) were released by washing the cellulose twice with 0.3 M NaCl, 50 mM Na phosphate pH 8, and lastly the HS lyase was removed by repeating the Ni-NTA purification. HS lyase digests the HS polysaccharide chains, leaving only the tetrasaccharide linkers (GlcAβ1-3Galβ1-3Galβ1-4Xylβ1) on the consensus serine residues of the HS attachment sites. Removal of HS chains from the wild type (GPC1-WT), and the proteins with disrupted N-glycosites, Gpc1-N79Q and Gpc1-N116Q, produced the Gpc1-WT-dHS, Gpc1-N79Q-dHS and Gpc1-N116Q-dHS proteins respectively.

Protein crystallization, dehydration and structure determination: Crystallization of Gpc1-dHS protein was performed using sitting drop vapour diffusion by mixing 2 µl of protein at 25 mg/ml with 2 µl of reservoir solution containing 14–16% PEG 6000, 0.1 M Tris-HCl pH 8.0 and 0.2 M $CaCl_2$, equilibrated over 0.5 ml of reservoir solution. Thin plate-like crystals of dimensions around 0.9 x 0.2 x 0.05 mm grew in 2 weeks within a lot of precipitated protein. Gpc1-dHS crystals were soaked in 16% PEG 6000, 0.2 M $CaCl_2$, 0.1 M Tris pH 8.0 and 15% ethylene glycol then mounted on a mesh LithoLoop (Molecular Dimensions, UK) in the HC1b machine at beamline I911-3 of the MAX IV Laboratory, Lund, Sweden. The Gpc1-dHS crystals were dehydrated for a total incubation time of 50-60 min to a final relative humidity (RHf) of 86–88%. The dehydrated crystals were then flash-frozen and stored in liquid nitrogen using the CATS sample changer (IRELEC, Saint-Martin-d-Hères, France), and subsequently complete diffraction data sets were collected at 100K. Diffraction images were indexed and scaled using XDS (15) and were further processed using programs from CCP4 package (16). The initial model was obtained by rigid body refinement of the dehydrated Gpc1-dC

model (PDB ID 4BWE) using REFMAC5 (17), followed by manual building in Coot (18) and rounds of refinement using phenix.refine (19). Finally, model validation was performed using Molprobity (20). PyMOL V1.6 (Schrödinger, LLC) was used for molecular rendering.

N-glycan characterization by HILIC-FLD-UPLC: The Gpc1 proteins (150 µg) were immobilized on a 10 kDa spin filter (Pall, Port Washington, NY), treated with denaturation buffer (50 µl of 50 mM DTT, 20 mM $NaHCO_3$) and incubated at 65 °C for 15 minutes to denature the proteins to allow for efficient cleavage of the N-glycans. After cooling to room temperature for 10 min, an iodoacetamide solution (50 mM, 50 µl per well) was added and incubated at room temperature for 30 min. Samples were spun at 12000 rpm for 5 minutes and the flow-through was discarded to waste, followed by two washings with milliQ water (50 µl) to remove residual amounts of denaturation reagents. N-glycans were enzymatically released by digestion with recombinant N-glycosidase F (50 µl, 0.5 mU in 20 mM $NaHCO_3$, pH 7.0) (Prozyme, Hayward, CA) and incubated at 37 °C overnight. After extraction, glycans were derivatized with 2-aminobenzamide (2-AB), sodium cyanoborohydride, 30 % v/v acetic acid in DMSO at 65 °C for 2 h. After cooling down to room temperature for 10 min, 100 µl of acetonitrile were added to the 2-AB labelled sample and loaded into a pre-conditioned Glycoworks HILIC cartridge (Waters, Milford, MA) for excess fluorophore removal. The cartridge was washed with 400 µl of 85 % of acetonitrile and glycans were eluted with 100 µl of 100 mM ammonium acetate in 5 % acetonitrile. The elution was repeated two more times and the samples were concentrated to dryness in a vacuum evaporator. The samples were dissolved in acetonitrile/ H_2O (70:30) and analyzed by HILIC-FLD-UPLC using a 1.7 µm BEH glycan column (2.1 mm x 15 mm, Waters, Milford, MA) and Waters ACQUITY UPLC I-class with fluorescence detection. The column temperature was kept at 40 °C and the flow rate set to 0.561 mL/min using a linear gradient of 50 mM ammonium formate (pH 4.4) against acetonitrile with ammonium formate from 30 % to 47 % over a 25 min period. An injection volume of 25 µl sample prepared in 70% v/v acetonitrile was used throughout. Fluorescence detection was achieved using excitation and emission wavelengths of 330 nm and 420 nm respectively.

A representative HILIC-FLD-UPLC profile of Gpc1 was annotated with glucose unit (GU) values by comparison with a dextran hydrosylate ladder (21). Initial structural assignment of the glycans present in the peaks was performed by comparison of experimental data with GlycoBase (http://glycobase.nibrt.ie) (22). The Consortium for Functional Glycomics (CFG) glycan notation is used throughout the manuscript (23) .

The sequence and linkage specificities of the glycans were analyzed by exoglycosidase enzymes purchased from Prozyme (Hayward, CA). 2-AB-labeled glycans were digested in 10 µl of supplied buffers at 37 °C overnight, using arrays of the following enzymes: ABS, *Arthrobacter ureafaciens* sialidase (EC 3.2.1.18, releases α2-3,6,8 linked non-reducing terminal sialic acid); NAN1, recombinant sialidase (EC 3.2.1.18, releases α2-3 linked non-reducing terminal sialic acids); BKF, bovine kidney α-fucosidase (EC3.2.1.51, releases α1-2,6 linked non-reducing terminal fucose residues more efficiently than α1-3,4 linked fucose, digests core α1-6 fucose); BTG, bovine testes β-galactosidase (EC 3.2.1.23, hydrolyses non-reducing terminal β1-4 and β1-3 linked galactose); GUH, hexosaminidase cloned from *Streptococcus pneumoniae* expressed in *E. coli* (EC 3.2.1.30, releases GlcNAc residues but not a bisecting GlcNAc linked to Man); and JBM, jack bean alpha mannosidase (EC 3.2.1.24, hydrolyses α(1-2,3,6)-linked mannose from oligosaccharides). After incubation, enzymes were removed by filtration using Pall spin filters 10kDa. N-Glycans were then analyzed by HILIC−FLD-UPLC as described above.

Analytical SEC and DLS: SEC analysis was performed at room temperature using a pre-equilibrated Superdex 200 10/300 GL column (GE Healthcare) in the standard buffers. Sample homogeneity was assessed by DLS using a Zetasizer APS DLS system (Malvern Instruments Ltd., Malvern, UK) under the same conditions as the SAXS measurements were done and the data were analyzed using Zetasizer software v7.03.

SAXS data collection and processing: Synchrotron radiation small-angle x-ray

scattering (SAXS) data were collected on the ID14-3 & BM29 beamlines, ESRF, Grenoble, France (24). Water measurements (empty capillary and water) were employed as a reference for further measurements and to give preliminary estimates for the sample molecular weight using the known absolute scattering of water ($I_{0,abs}$ (water)= $1.632.10^{-2}$ cm^{-1}, at 25 °C).
Size exclusion directly in-line with SAXS (SEC-SAXS) was used to obtain scattering data from highly pure monomers of Gpc1-dHS & Gpc1-dC using the HPLC system (Viscotek GPCmax, Malvern instruments) attached directly to the sample-inlet valve of the sample changer at BM29. 100 µl of ~6 mg/ml clarified sample was loaded onto a Superdex 200 10/300 GL column (GE Healthcare) pre-equilibrated with two column volumes of the protein standard buffer. The protein was eluted at flow rate of 0.5 ml/min, passed through the capillary cell and a scattering frame was collected every 2 s with a total of 1200 frames. The EDNA pipeline (25) provided a one-dimensional profile for each frame. The individual frame was subtracted and processed using tools from the ATSAS suite (26) calculating the forward scattering I(0) and radius of gyration R_g. Frames with stable R_g values from the monomer peak scattering intensity (the second half of the peak) were merged to provide a single averaged frame corresponding to the scattering of an individual SEC purified monomer.
For concentration series measurements, the samples were dialyzed overnight against the standard buffers and concentration series were prepared. A sample changer robot was used to load 35 µl of the clarified sample into the measurement capillary and SAXS data were collected in flow mode. Scattering profiles of the filtered dialysis buffers were subtracted from the corresponding sample scattering profiles. Real space R_g, excluded particle (Porod) volume and the maximum particle dimension (D_{max}) were estimated from the pair distance distribution function p(r) using the program GNOM. Molecular mass was determined using the program SAXSMoW. Porod-Debye plots were employed to analyse the protein flexibility according to Rambo & Tainer (27).

Ab initio modelling
Ab-initio shape reconstruction was performed using the simulated annealing method as implemented in the DAMMIN program (28). 20 different models were aligned, averaged and the most typical model was generated using DAMAVER program suite (29). To test the dependence of the SAXS shape reconstruction on methods, the program GASBOR (30) was also used to compute a set of 20 different *ab initio* shape envelopes using GNOM output up to 4.5 nm^{-1} but for the same D_{max} and R_g used with DAMMIN.

All-atom modelling, MES and EOM:
All-atom models (Allosmod) of Gpc1–dC & Gpc1-dHS were generated using tools on the SaliLab web server (http://salilab.org/index.html). ModLoop was used for modelling of loops missing in the Gpc1 crystal structure (PDB ID 4YWT) (31). No building was performed for the missing residues at the N- and C-termini of the core proteins. Next, static models of Gpc1-dC& Gpc1-dHS with flexible glycans were generated in MODELLER utilizing the ModLoop output. Monosialylated digalactosylated biantennary glycans (at position N79) and core-fucosylated monogalactosylated biantennary complex glycans with bisecting GlcNAc (at position N116), corresponding to the predominant glycoforms identified in the MS studies, were added with ideal geometries, followed by a 1 Å randomization of the all-atom coordinates, where the motions were restricted to loops and surface side chains. The generated core models were then used for further modelling jobs. Lastly, a 330 K AllosMod simulation was used to generate 2000 protein conformations with rigid sugars consistent with the input structure. The ensembles of glycosylated models were fitted to the raw SAXS data using FoXS (32). The χ^2 value was used to evaluate the goodness of fit and to select the AllosMod models of best/poorest-consistency with the SAXS data. Minimal ensemble searches (MES) with up to four conformations were tried to minimize the χ^2 values (33).
Furthermore, an ensemble containing a number of different conformations of the flexible parts was obtained by the EOM approach (34). The static Gpc1 model generated by MODELLER (residues 26 - 476) was used as a fixed core in RANCH to generate a large pool of 40 000 independent conformers of the N- & C-terminal residues of Gpc1-dHS with the four best orientations of N-glycans (obtained from Gpc1-dC MES results). GAJOE was used for ensemble selection by minimizing the discrepancy χ^2.

Rigid body modelling

Molecular modelling of the N- and C-termini was conducted using rigid body modelling as implemented in the program CORAL, based on the static model generated by MODELLER. N-glycan chains with structures identical to the ones used in Allosmod runs were treated as rigid bodies with contact restraints of 7 Å to the appropriate asparagine residues N79 & N116. Fifty independent CORAL runs were performed by minimizing the discrepancy χ^2 between the theoretical scattering curve calculated from the model and the experimental data. For Gpc1-dC dimer generation, 2-fold symmetry was applied. All CORAL models were aligned in Supcomb13 (35) and analyzed in PyMOL to identify the most typical structures. The theoretical scattering amplitude of the generated models together with the their discrepancies χ^2 to the experimental data, were calculated using CRYSOL (36).

Peptide mass spectrometry:

In-gel trypsin digestion of Gpc1 SDS-PAGE bands followed by LC-MS/MS analysis was performed to identify protein sequences as described previously (37). Data dependent mass spectrometry experiments were performed with an EASY LC Nano Flow high-performance liquid chromatography (HPLC) system (Proxeon Biosystems, Odense, Denmark) connected to an LTQ-Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, Waltham, WA) equipped with a nano Easy spray ion source (Proxeon Biosystems, Odense, Denmark). The chromatographic separation was performed at 40 °C on a 15 cm (75 µm i.d.) EASY-Spray column packed with 3µm resin (Proxeon Biosystems, Odense, Denmark). The nano HPLC intelligent flow control gradient was 5–20% solvent B (0.1% (v/v) FA, 100% (v/v) acetonitrile in water) in solvent A (0.1% (v/v) FA in water) for 120 min, then 20%–40% for 60 min followed by an increase to 90% for 5 min. A flow rate of 300 nl/min was used through the whole gradient. An MS scan (400–1400 m/z) was recorded in the Orbitrap mass analyzer set at a resolution of 60 000 at 400 m/z, 1×106 automatic gain control target and 500 ms maximum ion injection time. The MS was followed by data-dependent collision-induced dissociation MS/MS scans on the eight most intense multiply-charged ions in the LTQ at 500 signal threshold, 3 m/z isolation width, 10 ms activation time at 35 normalized collision energy and dynamic exclusion enabled for 60 seconds. The general mass spectrometric conditions were as follows: spray voltage 2.0 kV; no sheath or auxiliary gas flow; S-lens 60%; ion transfer tube temperature 275 °C. Raw data were processed by Mascot Distiller searching the SwissProt database (release 11-Dec-2013, containing 541954 entries) with an in-house Mascot database. The search parameters for the Mascot searches were: Taxonomy: *Homo sapiens,* Enzyme: trypsin or chymotrypsin, Variable Modifications: oxidation (Me), Precursor Tolerance: 20 ppm, MS/MS Fragment Tolerance: 0.1 Da.

RESULTS

We expressed and purified various 6xHis-tagged constructs of GPI anchorless Gpc1 (summarized in Table 1) as described previously (13,38). Briefly, we produced anchorless full length Gpc1 with and without HS chains (by site directed mutagenesis of HS attachment sites: S486A, S488A, S490A) (Gpc1 WT: 24-529 and Gpc1-dHS: 24-529), as well as C-terminally truncated Gpc1 lacking the C-terminal region (Gpc1-dC: 24-479). To investigate the structural roles of the individual N-glycans, two different Gpc1 constructs with disrupted N-glycosylation sites but intact HS attachment sites were also produced (Gpc1-N79Q: 24-529 & Gpc1-N116Q: 24-529).

Gpc1-dHS crystal dehydration and structure determination:

The full-length Gpc1-dHS protein crystallized in space group $P2_1$, containing 4 monomers in the asymmetric unit, with unit cell dimensions of a = 47.2, b = 169.0, c = 151.6 Å, β = 95.0°, and crystals diffracted to ~3 Å resolution (10). Similarly to crystals of Gpc1-dC, the Gpc1-dHS crystals were not isomorphous (with c dimension varying between 148–155 Å) and diffracted anisotropically, as revealed by a significantly higher B-factor in the c* dimension than in the a* & b* directions (ΔB = 56 Å^2) (39). Some parts of the structure, including the long C-terminus, were disordered and not visible in the electron-density map.

We have previously shown that controlled dehydration using the HC1b machine greatly improved the diffraction properties and amount of visible structure in Gpc1-dC crystals (12). Here, we investigated whether we could improve the diffraction quality of full-length Gpc1-dHS crystals using the same method, aiming to

resolve at least some of the structure of the C-terminal region. We optimized the dehydration protocol for Gpc1-dHS crystals and succeeded in reproducing isomorphous dehydrated crystals with unit cell dimensions of a = 46.8, b = 166.6, c = 137.7 Å, β = 90.4°, which diffracted to 2.1 Å in the best orientation. This represents a reduction in the c axis length by 13.9 Å. Complete 2.3 Å data were collected from a crystal dehydrated to a final relative humidity of 87 % with dehydration rate of 0.5% per 200 sec and total incubation time of 55 min in the humidified air stream of the HC1b machine (Table 2). The anisotropy ΔB (22 Å^2) and Wilson B factor (38.7 Å^2) of the new data were reduced by 61% and 27% respectively compared to non-dehydrated crystals. These revealed significant improvement in the lattice order and packing after dehydration, generated much better and less noisy electron density maps, allowing the building of more complete monomers in the asymmetric unit (5% more than in 4AD7) and displayed a better defined side-chain density. The final Gpc1-dHS model included residues P25–D475 with only a few missing residues (P350 to R360 & S408 to D412). The overall model B-factor fell from 74.8 to 59.1 Å^2 after dehydration. The backbone flexibility of Gpc1 as measured by the C_α B-factors was significantly higher than average for residues close to the N79 glycosylation site, the N-terminal region (including elements α1, L1, and part of α2) and parts of the protease site lobe (involving α4, α5, α11 and L2). The L3 loop was less flexible than the others, being stabilized by two disulphide bonds (Figure 1B). Unfortunately, no additional electron density was observed for the C-terminal domain after dehydration, which confirms that the HS attachment region is highly flexible and has no unique structure in the Gpc1-dHS crystal. This raises the question whether the C-terminus is intrinsically disordered or contains any locally ordered structure.

C-terminal flexibility and bioinformatics

Different intrinsic disorder predictors such as IUPred (40), PONDR-FIT (41), DISEBML (42), POODLE (43) and DISpro indicated that the C-terminus of Gpc1 (D475- T529) has a large tendency to be unstructured, with a slight propensity for higher order in the region from G493 to K506 (Figure 1C). Further, the secondary structure predictors PSIPRED (44) and JPRED (45) suggested the absence of secondary structure. However, the PROF (46), GLOBPLOT (47) and SSPro (48) predictors proposed a short low-complexity sequence between C494 & S507. Taken together, these analyses imply that indeed most of the Gpc1 C-terminus sequence lacks tertiary structure and that the HS attachment region and the last part of the C-terminus are the most likely regions to be structurally disordered as identified by all algorithms.

N-glycan characterization

The structures of the N-glycans decorating the Gpc1 core protein were investigated by chromatographic approaches. The total N-glycan pools of Gpc1-WT, Gpc1-N79Q and Gpc1-N116Q were assigned using exoglycosidase digestions and analysis on the 1.7 µm hydrophilic interaction liquid chromatography (HILIC) phase (Figure 2). The resulting data confirmed the structural diversity of the glycans in all the samples. Emphasis was placed on the identification of the most abundant peaks to use in SAXS modelling (see below) rather than a comprehensive analysis of the structures of all N-glycans. The Gpc1-WT profile consisted of 44 chromatographic peaks (Table S1). The largest proportions of oligosaccharides on Gpc1-WT have core-fucosylated biantennary structures with bisecting GlcNAc carrying one galactose (FA2BG1) and N-glycan core structure with 4 N-acetylhexosamine residues (HexNAc), which represent 9.7% of the total glycan pool. Core fucosylated biantennary structures with bisecting GlcNAc carrying two galactoses (FA2BG2), mannose-5 (M5), biantennary structure with bisecting GlcNAc carrying two galactoses and one sialic acid (A2BG2S1[6]), and lastly biantennary glycan with bisecting GlcNAc and one galactose (A2BG1) accounted for 5.8, 4.6%, 4.4%, and 4% of the total glycan pool, respectively (Table S2).

A series of exoglycosidase digestions was performed to identify and assign N-glycan structures to particular chromatographic peaks. The undigested profile of Gpc1-WT consisted of 44 chromatographic peaks, but the majority of the peaks had less than 2% abundance (Table S1). A panel of digestions for Gpc1 N-glycans that includes most commonly used exoglycosidase enzymes (α-sialidase, α-fucosidase, β-galactosidase, β-hexosaminidase and α-mannosidase) is shown in Figure 3. The digested glycans separated by HILIC have a logical movement in GU value, whereby each oligosaccharide residue can be accounted for by

a constant value, depending on linkage. The GU shifts followed the consecutive removal of terminal sugar residues. After sequential digestion with exoglycosidases we were able to assign the most abundant peaks.

After sialidase treatment (ABS digestion), biantennary structures with bisecting GlcNAc carrying two galactoses and one sialic acid (α2,6 linkage) with GU 8.30 digest back to biantennary structures with bisecting GlcNAc carrying two galactoses with GU 7.15. The digestion with NAN1 enzyme (a recombinant sialidase that removes α2,3-linked non-reducing terminal sialic acids) confirms that sialic acid is linked via an α2,6 linkage, as the addition of this enzyme does not move the peak. The peak with GU 7.65 does not move upon addition of sialidase enzymes but upon addition of α-fucosidase enzyme moves back by 0.5 GU, suggesting that is core-fucosylated. In addition treatment with β-galactosidase resulted in peak movement by 1.6 GU, corresponding to two galactose residues. The most abundant peak with GU 6.86 appeared to contain two co-eluting structures. Digestion with β-galactosidase confirmed presence of one galactose residue at the terminal end but some proportion (30%) of the peak remained after addition of sialidase, fucosidase, galactosidase and mannosidase, suggesting the presence of terminal HexNAc residues. Following enzymatic treatment with β-N-acetylhexosaminidase (GUH) the peak changed elution position. Unfortunately due to the specificity of the enzyme it is not possible to confirm whether terminal residue represents a GlcNAc or a GalNAc residue. Based on database matching, the structure could correspond to a tetra-antennary structure (A4) but also to the structure containing LacdiNAc residue (GalNAcβ1–4GlcNAc), which seems more likely since HEK cells were previously reported to express this structural feature (49,50).

The identification of GU 6.5 peak was primarily based on α-fucosidase digestion, as the peak did not change either its elution or migration position following enzymatic treatment, but its relative area increased due to digestion of core-fucosylated analogue. The presence of a high mannose structure in the peak with GU 6.17 was confirmed by database matching and also with α-mannosidase digestion.

In order to identify glycans originating from each glycosylation site, the Gpc1 mutants of Gpc1-N79Q and Gpc1-N116Q were analyzed by HILIC-FLD-UPLC, allowing for identification of the most abundant structures from the N116 and N79 sites respectively. Most of the structures overlapped between the two glycosylation sites; however, there were some trends specific for each one. Less complex glycans were more abundant at N116 (Gpc1-N79Q), whereas more complex, sialylated glycans were more prominent at N79 (Gpc1-N116Q) (Figure 2B & 2C).

Biophysical characterization of purified Gpc1

Analytical size-exclusion chromatography (SEC) of Gpc1-dC & Gpc1-dHS displayed a main elution peak representing a Gpc1 monomer (Figure 4A). The molecular weight of the Gpc1-dC monomer calculated from SEC (~62 kDa) was significantly smaller than that of the Gpc1-dHS monomer (~71 kDa), confirming the presence of the C-terminal domain in the dHS protein (verified by SDS-PAGE gel). The homogeneity of the eluted protein fractions was assessed by dynamic light scattering (DLS), without performing further concentration (Figure 4B). The volume distribution plots of the fractions from the second half of the SEC peak of Gpc1-dC & Gpc1-dHS showed single, monodisperse peaks with a polydispersity index (PI) of 11-14% and similar patterns with R_h = 4.1 nm for Gpc1-dC and 4.25 nm for Gpc1-dHS. An estimation of the molecular weight (MW) assuming a globular protein suggested a MW of 89.4 ± 12.6 kDa for Gpc1-dC and 99.6 ± 27 kDa for Gpc1-dHS. The reason for such molecular weight over-estimation can be the highly non-globular shape of Gpc1 and/or the flexibility of the N-glycan chains decorating the core protein.

Structural studies using SAXS

To elucidate the structure of glycosylated Gpc1 in solution, SAXS data were collected for both the glycosylated monomeric Gpc1-dC and the Gpc1-dHS proteins using inline size exclusion chromatography with SAXS (SEC-SAXS) (Figure 4 & Table 3). No sign of protein aggregation was observed. The SAXS curves of Gpc1-dC & Gpc1-dHS were distinctly similar over the whole q-range collected, except in the region between 0.8 and 1.5 nm^{-1}, producing similar R_g (~3.6 nm), but with a larger Porod volume for the full-length protein (about ~14% higher). The most accurate method for the estimation of the glycoproteins' MW was provided by SAXSMOW calculation (51), with an expected uncertainty of 10%. In our hands,

SAXSMOW estimated MWs of 67.0 kDa & 74.9 kDa for Gpc1-dC & Gpc1-dHS respectively, which are consistent with monomeric proteins. Thus, the solution scattering data meet the essential requirements for extracting accurate monomer shape information.

To estimate the distribution of masses within the particle and its shape, the pair-distance distribution P(r) function was calculated from the scattering data using GNOM (52). The P(r) profiles of Gpc1-dC & Gpc1-dHS showed main peaks around 2.75 & 2.87 nm respectively, with inclined distributions of vector lengths (Figure 4C). This indicates that both proteins have an extended structure in solution, with maximum dimensions of 118.0 & 119.5 Å respectively, consistent with the crystal structures. However, the P(r) plot of Gpc1-dHS exhibited an asymmetric fall-off accompanied by an extra shoulder around 5.8 nm that may be correlated to the C-terminal extension. The characteristic shape factor σ (the R_g/R_h ratio) for a globular protein is ~0.774; however when molecules deviate from globular shape to ellipsoidal, σ increases, as the R_g becomes larger than R_h (53). Gpc1-dC has σ = 0.9, while Gpc1-dHS has σ = 0.87, which reveals a more elongated structure for the truncated version of Gpc1 than for the full-length one. This is consistent with the idea that the C-terminal region may extend perpendicular to the protein surface (see later). To gain insight into Gpc1 flexibility, Porod Debye plots were calculated, and displayed a loss of the plateau in the full-length protein compared to the truncated version (Figure 4D). This implies that the scattering contrast had become more diffuse, suggesting increased flexibility in the presence of the C-terminal domain.

The DAMMIN ab-initio shape reconstruction program (28) was employed to generate 20 models of Gpc1-dC & Gpc1-dHS (Figure 4C insets & Table 3) with good structural convergence, as reflected in the low Normalized Spatial Discrepancy (NSD) values following structural alignment (NSD 0.71±0.02 and 0.64±0.02 for Gpc1-dHS and Gpc1-dC respectively; see SI Text). The resulting models revealed sizes and shapes consistent with the core protein substructures, in good agreement with the SAXS patterns (Figure 4E). Despite the fact that the N- & C-termini and two N-glycans should contribute to the SAXS patterns, they were not resolved as features in the averaged low-resolution envelopes. These would presumably provide extra protrusions to the envelopes close to their attachment sites in the individual bead models, which are averaged out during the process of averaging and filtering to generate the final conserved model. The Gpc1-dHS ab-initio model showed an additional small conserved bulge that protruded ~ 10 Å from the middle of the protein, proximal to the last C-terminal residue visible in the crystal structure (Figure 4C). To test the dependence of the ab-initio envelopes on the program used to determine them, a different algorithm, namely GASBOR, was also tested. Both GASBOR and DAMMIN produced truncated and full-length models with similar overall shapes (data not shown).

Structural reconstruction of the N-glycans decorating the Gpc1 core protein

To explore the N-linked glycan assembly on the Gpc1 core protein, we employed the Gpc1 crystal structure determined in this study to reconstruct the N-glycan structure using the SAXS data from GPC1-dC (i.e. lacking the C-terminal extension). Firstly, the structures of missing loops in Gpc1 were built and optimized using the ModLoop protocols (31). Subsequently, the predominant N-linked glycan structures and the N-terminus were reconstructed using all atom modelling (AllosMod; http://salilab.org/index.html, as described below. Modelling with the most abundant glycan structures has been demonstrated to give good agreement with SAXS data (54). The predominant N-glycan structures vary between different batches of purified protein, so we assigned the structures of the predominant glycoforms from the same protein batch that was used for SAXS data collection and these structures were used for further modelling. Monosialylated digalactosylated biantennary complex glycans (at position N79) and core-fucosylated monogalactosylated biantennary complex glycans with bisecting GlcNAc (at position N116) were the predominant glycoforms derived from the chromatographic data of purified Gpc1 without heparan sulphate (data not shown).

The simulated Gpc1-dC model without N-glycans gave a poor agreement with the observed SAXS data, even when the N-terminus was reconstituted by modelling (Figure 5A). In contrast, good agreement with the SAXS data was achieved with the glycosylated models, thus the sampling of glycan conformations

(accounting for ~9% of the total scattering mass) was crucial for generating accurate models. A comparison of the ten best- and poorest-fitting AllosMod models of Gpc1-dC indicated an obvious difference in the glycan orientation (Figure 5A & B). In the 10 best models (χ^2 = 1.0 ± 0.05), the two glycan chains protrude outward from helix α2 and the N-terminus spreads out close to the protein surface, whereas in the 10 worst fitting models (χ^2 = 3.99 ± 0.08), the glycan chains were localized close to the protein surface and the N-terminus protruded away from the core.

Fifty CORAL runs (26) were carried out to check the Gpc1-dC model reproducibility using a different algorithm than AllosMod. The proposed CORAL models were consistent throughout different modelling runs. The five CORAL models having the best agreement with the SAXS data (χ^2 = 0.9 ± 0.002) had the glycan chains consistently pointing outwards in similar orientations to the best models from Allosmod (data not shown). Nonetheless AllosMod has the advantage of an all-atom modelling approach to obtain stereochemically sound models.

Because of the inherent flexibility of the N-glycan chains, Minimal Ensemble Searches (MES) were attempted to improve the agreement with the SAXS data by including up to four models from each ensemble (33). MES provided a modest improvement in the consistency with the experimental SAXS data over the single model (χ^2 = 0.87 versus 0.94 for the best single model) (Figure 5A & C). Both of the N-glycans were modelled in proximity to high B-factor regions of the core protein crystal structure (Figure 1B) and oriented alongside helix α2 with higher locational variability for N79 glycans than for those on N116. The N79 glycan chains diverged widely over an area including the N-terminal helix α1, the hydrophobic part of L1 (GFSLSDVPQA) and the beginning of the L3 loop, suggesting that they are not involved in specific interactions with the core protein. In contrast, the N116 glycan chains were oriented in the vicinity of the end of L2 with smaller divergence (Figure 5C).

Spatial occupancy of the Gpc1 C-terminus in solution

To elucidate the structural orientation of the C-terminal residues and to allow an estimation of the distance between Gpc1 and the cell membrane, similar modelling strategies were used to build models that agree with the Gpc1-dHS SAXS data. The best N-glycosylated truncated Gpc1 structure lacking the C-terminal residues had poor agreement with the Gpc1-dHS SAXS data (χ^2 = 1.73, Figure 5D). Therefore, generating Gpc1-dHS SAXS consistent models require the capturing of the correct N- and C-terminals and glycan orientations. The C-terminus is not folded around the core protein in the best-fitting models (χ^2 = 0.960 ± 0.007; Figure 5E), but is rather extended towards the periphery, stretching for ~ 40 Å from the last helix in the crystal structure (α14). In the poorest-fitting models (χ^2 = 4.6 ± 0.1), the C-terminus was even more highly extended (> 70 Å). Further, the modelled positions of the N-terminus and N-glycan chains in the best and worst models generated by AllosMod were highly similar in both Gpc1-dC and Gpc1-dHS, which strengthens our confidence in the interpretation of these structural features. The CORAL approach reproduced the characteristic extension of the C-terminus obtained by the AllosMod method with comparable agreement with the experimental SAXS data (χ^2 = 0.92 ± 0.008; data not shown).

Because of the high number of variables in Gpc1-dHS modelling (N- & C-termini and two glycan chains), the probability of capturing a single conformation that most correctly explains the SAXS pattern is quite limited. MES was successful in improving the fit over a single model (χ^2 of 0.9 and 0.95 respectively). The best-fitting C–terminal conformations form a compact cluster, spreading between 35 to 40 Å in a direction perpendicular to the core protein, and mainly localized between the central and the protease site lobes (Figure 5F). Furthermore, EOM modelling (34) was used to check the consistency of the MES modelling processes. The EOM ensemble of models of Gpc1-dHS showed the C-terminus in the same orientation as in the MES results, but with a shorter extension (between 30–35 Å; data not shown). The Gpc1 membrane proximal surface, defined by this orientation of the C-terminus, is comprised of the conserved α14, L1 & L3 in the Cys-rich lobe and α4 & α5 in the protease-site lobe (Figure 10). The N79 glycan chains were directed perpendicular to the C-terminus while the N116 glycans were located on the opposite surface of the protein.

Does Gpc1 self-associate in solution?

In the SEC experiments, when we increased the loading concentration of Gpc1-dC to ≥ 1.5

mg/ml, the position of the main peak moved towards a higher mass and a small shoulder was eluted at 12.4 ml (Figure 6A), suggesting a protein dimer. This shift in the eluted peaks might originate from various monomer-dimer mixtures in the solution. SAXS measurements of Gpc1-dC at different concentrations (0.75-6 mg/ml) showed apparent changes in the low-q region indicative of concentration-dependent self-association or oligomerization (Figure 6B & Table 4). A small change in the SAXS curves (particularly at 6 mg/ml) beyond the Guinier region was observed when compared with Gpc1-dC SEC-SAXS data, which supports the emergence of dimer species at high concentrations. Furthermore, the measured R_h of GPC1-dC from DLS was directly proportional to the protein concentration, with a dramatic difference between the monomer (collected from the SEC fractions) and concentrated samples (Table 5). Importantly, the intermediate samples showed high polydispersity index, suggesting the presence of multiple species in the samples.

The DLS data show that at 6 mg/ml concentration, Gpc1-dC forms a highly extended dimer with a shape factor of 1.2. Moreover its SAXS P(r) plot showed a main peak at 2.7 nm with an extra shoulder at 5.9 nm, introduced by the dimerization (Figure 6B inset). Hybrid modelling by DAMMIN and CORAL (26) of the SAXS data collected at 6 mg/ml suggested an elongated architecture for the dimer in which the N-terminus was predominantly extended away from the protein (Figure 7). Gpc1-dC dimerized in the vicinity of the protease site lobe, particularly the region of α5 & α6, which is close to the location of the C-terminus in the Gpc1-dHS SAXS models (Figure 7). In contrast, little SEC peak shift or change in R_h was observed for the Gpc1-dHS samples (Figure 6A). Furthermore, all SAXS patterns collected at different Gpc1-dHS concentrations have similar shapes, with little difference in the Guinier region and P(r) when compared to the Gpc1-dHS monomer SAXS curve (Figure 6C). Taken together, this indicates that self-association does not occur for Gpc1-dHS in the tested concentration range 0.75 to 6 mg/ml and suggests that the presence of the C-terminal tail prevents unwanted interactions of the core proteins.

N-glycosylation protects aggregation- and degradation-prone regions of Gpc1

EndoH enzymatic removal of the N-glycans from Gpc1-dHS was achieved as described in the Methods section, producing the deglycosylated Gpc1-dHS-dN with a MW ~5 kDa lower than that of the glycosylated protein, as shown on an SDS-PAGE gel (Figure 8A). Importantly, SAXS measurements of Gpc1-dHS-dN showed little concentration-dependent behaviour, displaying invariability in the q-range beyond the Guinier region, which speaks against variation in the oligomeric structure in the concentration range 0.7–2.5 mg/ml. However, models generated by DAMMIN & CORAL using the merged SAXS data of dHS-dN revealed an elongated dimeric structure rather than a monomer (Figure 8B & C). Gpc1-dHS-dN dimerized through parts of the Cys-rich lobe that are covered by the N79 glycans in the glycosylated Gpc1. Removing the glycans from Gpc1-dC resulted in complete aggregation during purification. Furthermore, the mutant N79Q showed a high propensity for aggregation after removing the HS chains (Gpc1-N79Q-dHS; data not shown). Taken together, these data suggest a vital role for N79 glycans in preventing protein aggregation, at least in vitro. The solubility difference between Gpc1-N79Q-dHS & deglycosylated Gpc1-dHS might originate from the first GlcNAc residue linked to N79, which is not removed by EndoH. This GlcNAc confers most of the thermal stabilization effect of the N-glycans on Gpc1, as reported previously (13), and it seems likely that it also confers solubility enhancement.

In addition to aggregation, the Gpc1-N79Q samples showed degradation on SDS PAGE gels, with a strong extra band at ~47 kDa (Figure 9A), which indicates that the lack of the N79-glycosylation has rendered N79Q susceptible to intra- and/or extracellular proteolytic degradation. This proteolysis site was identified by a mass spectrometric analysis, and shown to be close to the beginning of the L3 loop (between K404-R414). This evidence reveals a role for N79 glycans in protecting against degradation of Gpc1.

We used highly monodisperse preparations of purified Gpc1-WT-dHS and Gpc1-N116Q-dHS (after removal of the HS by HS lyase) for SAXS data collection at different protein concentrations (up to 3 mg/ml) (Figure 9A &B). The HS lyase enzyme degraded the HS chains but did not remove the tetrasaccharide linker attached to the serine. The Gpc1-WT-dHS and Gpc1-N116Q-dHS scattering profiles were smoother than for

the proteins without the HS linkers and had almost identical concave patterns accompanied by negligible differences in R_g and D_{max} (Table 4), without concentration-dependent self-association or aggregation at the tested concentrations. Their MW and Porod volumes were highly overestimated, presumably because of the flexibility of the partially glycosylated C-terminal tails (bound to three tetrasaccharide linkers) and/or the presence of some oligomers. The pair distribution functions of Gpc1-WT-dHS & Gpc1-N116Q-dHS suggested similar shapes, with Gpc1-WT-dHS being slightly larger due to its extra N116-linked glycan chain (Figure 9C). The generated ab-initio envelopes featured similar overall expanded shapes, with larger volume for the WT samples. Attempts to model Gpc1-WT-dHS & Gpc1-N116Q-dHS using AllosMod and CORAL failed due to the complexity at the C-terminus, with three flexible tetrasaccharide chains in very confined region, thus the programs were unable to model them correctly. We concluded from comparison of Gpc1-WT-dHS & Gpc1-N116Q-dHS that no structural changes in Gpc1 (at the resolution level of SAXS) could be noticed after disrupting the N116 site, and therefore that this glycan might have other functional roles.

DISCUSSION

Glypicans are multifunctional GAG-substituted proteoglycans involved in the regulation of several cellular signalling pathways. Defects in their function lead to developmental distortions (2,55). The regulatory activity of Gpcs is based on their ability to either inhibit or stimulate the interaction of many growth factors with their signalling receptors. GAG chains are responsible for many of the biological functions of Gpcs. However, recent studies suggest regulatory roles for Gpc core proteins, for example in mediating cell signalling by direct binding to e.g. BMP4, FGF2, Wnt and Hedgehog (56-58).

The crystal structure of the Gpc1 core protein reveals a quite rigid, elongated single-domain α-helical fold, although flexibility is higher towards the ends. Almost identical disulphide connectivity patterns and structural similarity between the Cys-rich lobe of Gpc1 and the Cys-rich domain of the Frizzled receptors (functioning in Wnt signalling) has been reported (59), but whether this has functional relevance is still unknown.

Structural features of the Gpc1 N-glycans

The long α2 helix of the Gpc1 structure traverses the entire length of the protein and carries two N-glycan chains, one at each end (10). The Gpc1 N-glycan chains are not fully resolved in the crystal structure, due to intrinsic heterogeneity and flexibility. Glycoproteins are usually micro-heterogeneous, presenting a range of different N-glycans at each glycosylation site. The expression system and/or the folded protein itself can influence the N-glycan diversity, probably by affecting the substrate availability and the proximity of the N-glycosylation sites to the Golgi glycosyltransferases and glycosidases (60).

The N-glycan analyses presented here have revealed highly heterogeneous complex-type glycoforms for Gpc1, but it is not clear whether this heterogeneity has a functional impact. Characteristics of the N79-linked glycans include a high incidence of core fucosylation and sialylation, presenting relatively strong electronegative charges, which we discovered, based on the SAXS modelling, to cover hydrophobic patches on the Cys-rich lobe, including parts of the L1 loop and the α1 and α2 helices, that would otherwise promote aggregation. Furthermore, these N79 oligosaccharides seem to limit intra- and/or extra-cellular proteolysis of the Gpc1 core protein, specifically at the beginning of the L3 loop, extending the protein stability and lifetime, whereas removal of N116 glycans does not affect the overall structural characteristics of Gpc1.

Topology of the Gpc1 core protein, N- & O-glycans with respect to the cell surface

The Gpc1 C-terminal region lacks significant secondary or tertiary structure and is disordered in Gpc1-dHS crystals, even after controlled crystal dehydration, which otherwise significantly improves crystalline order, resolution and diffraction anisotropy. In view of this, we coupled our crystallographic knowledge on Gpc1 with SAXS to obtain information on the localization of the C-terminus and thus the potential spatial orientations of Gpc1 relative to the cell surface. Our results show that the C-terminal domain is highly flexible and extends approximately 35–40 Å from the core protein. Molecular dynamics simulations have shown that the GPI anchor of many glypiated proteins is highly flexible, but that it nevertheless can retain the proteins at distances between 9–13 Å from the cell surface, with minor impact on the

proteins' degrees of freedom for movement and orientation relative to the membrane (61,62). Accordingly, adding the length of the GPI anchor to that of the C-terminal tail, we suggest that the Gpc1 core protein is located approximately ~44–53 Å from the cell membrane. This distance is likely sufficient for HS assembly enzymes or even a membrane receptor candidate to interact with the membrane proximal surface of Gpc1. It seems likely that Gpc1 "lies down" in a transverse orientation to the membrane, with the Gpc1 orthologues' evolutionarily conserved surface (L1 and L3 loops, α4, α5 and α14 helices) facing the cell surface. No evolutionary conservation has been detected for the other surfaces of Gpc1 (Figure 10). Evolutionary conservation of these surface-exposed residues suggests that they are implicated in interaction with other macromolecules related to Gpc1 function. In the *Drosophila* glypican Dally-like core protein, structure-guided mutational evidence suggested that helices α4 and α5 are important for mediating Hedgehog signalling (lacking the HS chains) (63), consistent with these being oriented towards the membrane. The flexibility of the C-terminal region presumably allows great freedom for Gpc1 and other glypicans to reorient to accommodate binding to receptors and other signalling molecules, generally with the participation of the HS chains.

Bioinformatics tools predict a highly disordered structure for the anionic linker (10 residues) between the folded core protein and the GAG attachment site (at least 10–15 Å). Consequently, we can hypothesize that the HS attachment sites are located more than 30 Å from the membrane, and therefore could mediate the interaction of glypicans with other cell surface proteins. Previous work has reported that the HS chains do not stabilize, and probably do not interact with, the Gpc1 core protein (38), which is consistent with the extended C-terminal structure that we observe. However, the HS chains on Gpc1 protect the protein against irreversible aggregation, probably by providing the protein additional negative charge, resulting in electrostatic repulsion.

Hence we predict that the anionic HS chains extend from the C-terminal region in a plane initially approximately parallel to the membrane, then more divergent from it, to avoid repulsive interactions with the negatively charged phospholipid bilayer (Figure 10). Glypican HS chains can contain 50-150 repeated disaccharides. Recent work indicates that short 12-disaccharide HS polymers are extended in solution, with a partly bent conformation, up to 10 nm in length (64). Therefore, in mature Gpc1, the HS chains could be extended to 3–10 times as long as the compact cylinder of the core protein (40-130 nm). Therefore, Gpc1 could bind to the signalling molecules directly via its core protein, through its extended HS side chains or both. To what extent the HS chains might contribute to the conformation or stability of the C-terminal region requires further investigation.

The N116-linked oligosaccharides protrude from the membrane-distal surface of the core protein, which limits their potential interactions with the lipid bilayer and the HS. In contrast, the bulky negatively charged N79-glycans are directed perpendicularly to the evolutionarily conserved membrane-proximal Gpc1 surface (Figure 10).

The membrane proximal surface on Gpc1 and HS assembly

Glypiated proteins are often associated into ordered microdomains of the membrane called lipid rafts. However, it has been shown that the GPI anchor does not target the glypiated protein to the apical surface of epithelial cells, while the N-glycans do (65). Thus it is reasonable to suggest that the evolutionarily conserved surface of the core protein and/or the N79-glycans have a dual role: they assist both in targeting and associating the Gpc1 protein to the consensus lipid raft domains of the cell surface and in GAG substitution on the C-terminal domain.

Gpc1 is exclusively substituted with HS, whereas other glypicans like Gpc5 possess both HS and CS (66). Chen et al. (2001) concluded that the core protein plays a vital role in directing the assembly of HS rather than CS on rat Gpc1, as expression of the GAG attachment domain without the core protein results in substitution with ~90% CS. Moreover, mutational analysis shows that the L3 loop, α14 and other nearby parts of the core protein are required for preferential HS assembly (67). Previous work and the present data both confirm that the N-glycosylations of Gpc1 are not involved in GAG class determination but do affect the amount of HS substitution and also chain elongation (13). Accordingly, parts or all of the surface conserved elements (L1, L3, α4, α5 and α14) would be predicted to be involved in GAG class determination and synthesis by interacting with some components of the GAGs

biosynthetic pathway e.g. glycosyltransferases, necessary cofactors for the enzyme activity or some components required for trafficking the protein within the Golgi apparatus to regulate the HS substitution on Gpc1. HS biosynthesis is mediated by Golgi apparatus transmembrane glycosyltransferases of the exostosin family (EXTs), which initiate, elongate and terminate HS backbone formation (68). Five members have been identified in mammals, including EXT1, EXT2, EXTL1, EXTL2 and EXTL3. In general, the published results so far suggest that EXTL3 works as an initiator of HS chain biosynthesis, as no HS was detected in 9-day old mouse embryos lacking EXTL3 (69). Future work would be of importance to determine precisely how the Gpc1 core protein regulates HS class determination and assembly. This could be by searching for the Gpc1 interacting partner from the exostosin family and in particular to study its affinity and interaction with the EXTL3 as HS-initiator enzyme. Furthermore, systematic mutagenesis studies of the evolutionarily conserved surface structural elements would be helpful to map more distinctly the residues of the Gpc1 core protein that might be involved in the preferential HS assembly.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest with the content of this article.

AUTHOR CONTRIBUTIONS

WA purified all glypican constructs, crystallised them, performed crystal dehydration, collected data and refined the crystal structure, conceived and carried out SAXS experiments. BA, JÖ and NK carried out the analysis of the glycosylation patterns on Gpc-1. KM expressed the glypican constructs. DTL and KM conceived and supervised the project. All authors contributed to writing the paper.

REFERENCES

1. Brown, D., and Waneck, G. L. (1992) Glycosyl-phosphatidylinositol-anchored membrane proteins. *Journal of the American Society of Nephrology* **3**, 895-906
2. Fico, A., Maina, F., and Dono, R. (2011) Fine-tuning of cell signaling by glypicans. *Cellular and molecular life sciences : CMLS* **68**, 923-929
3. Whitelock, J. M., and Iozzo, R. V. (2005) Heparan Sulfate: A Complex Polymer Charged with Biological Activity. *Chemical Reviews* **105**, 2745-2764
4. Awad, W., Logan, D. T., and Mani, K. (2014) GPC1 (glypican 1). *Atlas Genet Cytogenet Oncol Haematol* **18**, 461-464
5. Jen, Y. H., Musacchio, M., and Lander, A. D. (2009) Glypican-1 controls brain size through regulation of fibroblast growth factor signaling in early neurogenesis. *Neural development* **4**, 33
6. Hooper, N. M. (2011) Glypican-1 facilitates prion conversion in lipid rafts. *Journal of neurochemistry* **116**, 721-725
7. Watanabe, N., Araki, W., Chui, D. H., Makifuchi, T., Ihara, Y., and Tabira, T. (2004) Glypican-1 as an A beta binding HSPG in the human brain: Its localization in DIG domains and possible roles in the pathogenesis of Alzheimer's disease. *Faseb Journal* **18**, 1013-+
8. Matsuda, K., Maruyama, H., Guo, F., Kleeff, J., Itakura, J., Matsumoto, Y., Lander, A. D., and Korc, M. (2001) Glypican-1 is overexpressed in human breast cancer and modulates the mitogenic effects of multiple heparin-binding growth factors in breast cancer cells. *Cancer research* **61**, 5562-5569
9. Su, G., Meyer, K., Nandini, C. D., Qiao, D., Salamat, S., and Friedl, A. (2006) Glypican-1 is frequently overexpressed in human gliomas and enhances FGF-2 signaling in glioma cells. *The American journal of pathology* **168**, 2014-2026
10. Svensson, G., Awad, W., Hakansson, M., Mani, K., and Logan, D. T. (2012) Crystal structure of N-glycosylated human glypican-1 core protein: structure of two loops evolutionarily conserved in vertebrate glypican-1. *The Journal of biological chemistry* **287**, 14040-14051

11. Bowler, M. W., Mueller, U., Weiss, M. S., Sanchez-Weatherby, J., Sorensen, T. L. M., Thunnissen, M. M. G. M., Ursby, T., Gobbo, A., Russi, S., Bowler, M. G., Brockhauser, S., Svensson, O., and Cipriani, F. (2015) Automation and Experience of Controlled Crystal Dehydration: Results from the European Synchrotron HC1 Collaboration. *Crystal Growth & Design* **15**, 1043-1054
12. Awad, W., Svensson Birkedal, G., Thunnissen, M. M., Mani, K., and Logan, D. T. (2013) Improvements in the order, isotropy and electron density of glypican-1 crystals by controlled dehydration. *Acta crystallographica. Section D, Biological crystallography* **69**, 2524-2533
13. Svensson, G., Hyrenius Wittsten, A., Linse, S., and Mani, K. (2011) The structural role of N-linked glycans on human glypican-1. *Biochemistry* **50**, 9377-9387
14. Chang, V. T., Crispin, M., Aricescu, A. R., Harvey, D. J., Nettleship, J. E., Fennelly, J. A., Yu, C., Boles, K. S., Evans, E. J., Stuart, D. I., Dwek, R. A., Jones, E. Y., Owens, R. J., and Davis, S. J. (2007) Glycoprotein structural genomics: solving the glycosylation problem. *Structure* **15**, 267-273
15. Kabsch, W. (2010) Integration, scaling, space-group assignment and post-refinement. *Acta crystallographica. Section D, Biological crystallography* **66**, 133-144
16. Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., and Wilson, K. S. (2011) Overview of the CCP4 suite and current developments. *Acta crystallographica. Section D, Biological crystallography* **67**, 235-242
17. Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta crystallographica. Section D, Biological crystallography* **67**, 355-367
18. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography* **66**, 486-501
19. Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., and Adams, P. D. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta crystallographica. Section D, Biological crystallography* **68**, 352-367
20. Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica. Section D, Biological crystallography* **66**, 12-21
21. Royle, L., Radcliffe, C. M., Dwek, R. A., and Rudd, P. M. (2006) Detailed Structural Anlysis of N-glycans Released From Glycoproteins in SDS-PAGE Gel Bands Using HPLC Combined With Exoglycosidase Array Digestions. in *Methods in Molecular Biology* (I.Brockhausen ed.), Humana Press Inc., Totowa, NJ. pp 125-143
22. Campbell, M. P., Royle, L., Radcliffe, C. M., Dwek, R. A., and Rudd, P. M. (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* **24**, 1214-1216
23. Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Marth, J. D., Bertozzi, C. R., Hart, G. W., and Etzler, M. E. (2009) Symbol nomenclature for glycan representation. *PROTEOMICS* **9**, 5398-5399
24. Pernot, P., Round, A., Barrett, R., De Maria Antolinos, A., Gobbo, A., Gordon, E., Huet, J., Kieffer, J., Lentini, M., Mattenet, M., Morawe, C., Mueller-Dieckmann, C., Ohlsson, S., Schmid, W., Surr, J., Theveneau, P., Zerrad, L., and McSweeney, S. (2013) Upgraded ESRF

BM29 beamline for SAXS on macromolecules in solution. *Journal of Synchrotron Radiation* **20**, 660-664
25. Incardona, M. F., Bourenkov, G. P., Levik, K., Pieritz, R. A., Popov, A. N., and Svensson, O. (2009) EDNA: a framework for plugin-based applications applied to X-ray experiment online data analysis. *J Synchrotron Radiat* **16**, 872-879
26. Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V., and Svergun, D. I. (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *Journal of Applied Crystallography* **45**, 342-350
27. Rambo, R. P., and Tainer, J. A. (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* **95**, 559-571
28. Svergun, D. I. (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing (vol 76, pg 2879, 1999). *Biophysical Journal* **77**, 2896-2896
29. Volkov, V. V., and Svergun, D. I. (2003) Uniqueness of ab initio shape determination in small-angle scattering. *Journal of Applied Crystallography* **36**, 860-864
30. Svergun, D. I., Petoukhov, M. V., and Koch, M. H. J. (2001) Determination of domain structure of proteins from X-ray solution scattering. *Biophysical Journal* **80**, 2946-2953
31. Fiser, A., and Sali, A. (2003) ModLoop: automated modeling of loops in protein structures. *Bioinformatics* **19**, 2500-2501
32. Schneidman-Duhovny, D., Hammel, M., and Sali, A. (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic acids research* **38**, W540-W544
33. Pelikan, M., Hura, G. L., and Hammel, M. (2009) Structure and flexibility within proteins as identified through small angle X-ray scattering. *General physiology and biophysics* **28**, 174-189
34. Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M., and Svergun, D. I. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *Journal of the American Chemical Society* **129**, 5656-5664
35. Kozin, M. B., and Svergun, D. I. (2001) Automated matching of high- and low-resolution structural models. *Journal of Applied Crystallography* **34**, 33-41
36. Svergun, D., Barberato, C., and Koch, M. H. J. (1995) CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography* **28**, 768-773
37. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2007) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protocols* **1**, 2856-2860
38. Svensson, G., Linse, S., and Mani, K. (2009) Chemical and thermal unfolding of glypican-1: protective effect of heparan sulfate against heat-induced irreversible aggregation. *Biochemistry* **48**, 9994-10004
39. Strong, M., Sawaya, M. R., Wang, S., Phillips, M., Cascio, D., and Eisenberg, D. (2006) Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences* **103**, 8060-8065
40. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433-3434
41. Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K., and Uversky, V. N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochimica et biophysica acta* **1804**, 996-1010

42. Linding, R., Jensen, L., Diella, F., Bork, P., Gibson, T., and Russell, R. (2003) Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453 - 1459
43. Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y., and Noguchi, T. (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* **23**, 2046-2053
44. Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K., and Jones, D. T. (2013) Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic acids research* **41**, W349-357
45. Cole, C., Barber, J. D., and Barton, G. J. (2008) The Jpred 3 secondary structure prediction server. *Nucleic acids research* **36**, W197-W201
46. Ouali, M., and King, R. D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein science : a publication of the Protein Society* **9**, 1162-1176
47. Linding, R., Russell, R., Neduva, V., and Gibson, T. (2003) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic acids research* **31**, 3701 - 3708
48. Cheng, J., Randall, A., Sweredoski, M., and Baldi, P. (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic acids research* **33**, W72 - W76
49. Hamouda, H., Kaup, M., Ullah, M., Berger, M., Sandig, V., Tauber, R., and Blanchard, V. (2014) Rapid Analysis of Cell Surface N-Glycosylation from Living Cells Using Mass Spectrometry. *J Proteome Res* **13**, 6144-6151
50. Kim, H. I., Saldova, R., Park, J. H., Lee, Y. H., Harvey, D. J., Wormald, M. R., Wynne, K., Elia, G., Kim, H.-J., Rudd, P. M., and Lee, S.-T. (2013) The Presence of Outer Arm Fucose Residues on the N-Glycans of Tissue Inhibitor of Metalloproteinases-1 Reduces Its Activity. *J Proteome Res* **12**, 3547-3560
51. Fischer, H., de Oliveira Neto, M., Napolitano, H. B., Polikarpov, I., and Craievich, A. F. (2010) Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *Journal of Applied Crystallography* **43**, 101-109
52. Svergun, D. I. (1992) Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. *Journal of Applied Crystallography* **25**, 495-503
53. Brewer, A. K., and Striegel, A. M. (2011) Characterizing the size, shape, and compactness of a polydisperse prolate ellipsoidal particle via quadruple-detector hydrodynamic chromatography. *The Analyst* **136**, 515-519
54. Guttman, M., Weinkam, P., Sali, A., and Lee, K. K. (2013) All-Atom Ensemble Modeling to Analyze Small-Angle X-Ray Scattering of Glycosylated Proteins. *Structure (London, England : 1993)* **21**, 321-331
55. Filmus, J., Capurro, M., and Rast, J. (2008) Glypicans. *Genome biology* **9**, 224
56. Kirkpatrick, C. A., Knox, S. M., Staatz, W. D., Fox, B., Lercher, D. M., and Selleck, S. B. (2006) The function of a Drosophila glypican does not depend entirely on heparan sulfate modification. *Developmental biology* **300**, 570-582
57. Cheng, W., Tseng, C.-J., Lin, T. T. C., Cheng, I., Pan, H.-W., Hsu, H.-C., and Lee, Y.-M. (2008) Glypican-3-mediated oncogenesis involves the Insulin-like growth factor-signaling pathway. *Carcinogenesis* **29**, 1319-1326
58. Williams, E. H., Pappano, W. N., Saunders, A. M., Kim, M.-S., Leahy, D. J., and Beachy, P. A. (2010) Dally-like core protein and its mammalian homologues mediate stimulatory and inhibitory effects on Hedgehog signal response. *Proceedings of the National Academy of Sciences* **107**, 5869-5874
59. Pei, J., and Grishin, N. V. (2012) Cysteine-rich domains related to Frizzled receptors and Hedgehog-interacting proteins. *Protein science : a publication of the Protein Society* **21**, 1172-1184

60. Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E. (2009) *Essentials of Glycobiology.*, 2nd ed.,
61. Zuegg, J., and Gready, J. E. (2000) Molecular dynamics simulation of human prion protein including both N-linked oligosaccharides and the GPI anchor. *Glycobiology* **10**, 959-974
62. Rudd, P. M., Morgan, B. P., Wormald, M. R., Harvey, D. J., van den Berg, C. W., Davis, S. J., Ferguson, M. A. J., and Dwek, R. A. (1997) The Glycosylation of the Complement Regulatory Protein, Human Erythrocyte CD59. *Journal of Biological Chemistry* **272**, 7229-7244
63. Kim, M.-S., Saunders, A. M., Hamaoka, B. Y., Beachy, P. A., and Leahy, D. J. (2011) Structure of the protein core of the glypican Dally-like and localization of a region important for hedgehog signaling. *Proceedings of the National Academy of Sciences* **108**, 13112-13117
64. Khan, S., Fung, K. W., Rodriguez, E., Patel, R., Gor, J., Mulloy, B., and Perkins, S. J. (2013) The solution structure of heparan sulfate differs from that of heparin: implications for function. *The Journal of biological chemistry* **288**, 27737-27751
65. Pang, S., Urquhart, P., and Hooper, N. M. (2004) N-Glycans, not the GPI anchor, mediate the apical targeting of a naturally glycosylated, GPI-anchored protein in polarised epithelial cells. *Journal of Cell Science* **117**, 5079-5086
66. Li, F., Shi, W., Capurro, M., and Filmus, J. (2011) Glypican-5 stimulates rhabdomyosarcoma cell proliferation by activating Hedgehog signaling. *The Journal of cell biology* **192**, 691-704
67. Chen, R. L., and Lander, A. D. (2001) Mechanisms underlying preferential assembly of heparan sulfate on glypican-1. *The Journal of biological chemistry* **276**, 7507-7517
68. Busse-Wicher, M., Wicher, K. B., and Kusche-Gullberg, M. (2014) The extostosin family: Proteins with many functions. *Matrix Biology* **35**, 25-33
69. Takahashi, I., Noguchi, N., Nata, K., Yamada, S., Kaneiwa, T., Mizumoto, S., Ikeda, T., Sugihara, K., Asano, M., Yoshikawa, T., Yamauchi, A., Shervani, N. J., Uruno, A., Kato, I., Unno, M., Sugahara, K., Takasawa, S., Okamoto, H., and Sugawara, A. (2009) Important role of heparan sulfate in postnatal islet growth and insulin secretion. *Biochemical and biophysical research communications* **383**, 113-118
70. Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic acids research* **38**, W529-W533

FOOTNOTES

Data deposition. The structure has been deposited in the Protein Data Bank, www.pdb.org (PDB ID code: 4YWT)

Acknowledgments. We wish to thank staff at beamline I911-3 of the MAX IV Laboratory for help with data collection, in particular Marjolein Thunnissen for assistance with the HC1b machine. We thank staff at the ESRF SAXS beamlines (Adam Round and Martha Brennich) for help with SAXS data collection. We also appreciate the assistance of Sven Kjellstrom with peptide MS. This work was supported by grants from the Swedish Research Council, European Union FP7 GastricGlycoExplorer ITN (316929), the Swedish Cancer Society, Alfred Österlund, Åhlen and Olle Engkvist Foundation. The HPLC instrument was obtained using a gift from Ingabritt and Arne Lundbergs Research Foundation.

FIGURE LEGENDS

Figure 1 Bioinformatics & crystal structure of Gpc1: (A) A schematic of human Gpc1 core protein linked to the cell membrane by its GPI anchor, with N- & O- glycosylation positions annotated and the Cys residues displayed as yellow lines. (B) Schematic representation of the Gpc1-dHS crystal structure determined in this work (PDB entry 4YWT) colored by B-factor distribution, where dark blue=20 A^2 and red=120A^2. (C) Prediction of the C-terminus disorder tendency using different algorithms. The sequences with a slight propensity for ordered structure are in bold face and their predicted secondary structures are shown below.

Figure 2 Overall analyses of N-glycans from purified Gpc1. The N-glycans present on Gpc1-WT (A), Gpc1-N79Q (B) and Gpc1-N116Q (C) as determined using exoglycosidase digestions and analysis on the 1.7 µm HILIC phase are shown. For each sample, the most abundant *N*-glycans are labelled and percentage area of the total profile, glucose unit (GU) and retention time (RT) are shown in Table S2. Nomenclature from (23).

Figure 3 HILIC-FLD-UPLC profiles of Gpc1 N-glycans subjected to exoglycosidase digestions in order to assign their structures. UND represents the undigested pool of glycans; this pool was digested with sialidase (ABS and NAN1), α-fucosidase (BKF), β-galactosidase (BTG), α-mannosidase (JBM), hexosaminidase (GUH) and a combination of enzymes.

Figure 4 Size characterization and SAXS study of Gpc1-dC (red) & Gpc1-dHS (blue): (A) SEC-SAXS profiles and the calculated R_g (faint red for Gpc1-dC & faint blue for Gpc1-dHS) obtained for each frame of the eluted proteins. (B) Normalized DLS size distribution plot with annotated R_h. (C) SAXS-derived normalized P(r) plots from Gpc1 samples in solution. Ab-initio averaged models of Gpc1-dC and Gpc1-dHS derived from the P(r) data are depicted as surface volumes and their contributions to the major and the minor peaks are indicated. (D) Porod-Debye plots, which indicate higher flexibility for the Gpc1-dHS protein than Gpc1-dC. (E) SAXS profiles with the fit for the DAMMIN ab-initio model.

Figure 5 Molecular modelling of Gpc-1 N-glycans and the N- & C-termini by SAXS: (A) Comparison of experimental scattering data of the Gpc1-dC protein (grey dots) and the fit of the deglycosylated model (green), the AllosMod best models (red), worst models (blue) & MES models (orange) respectively. (D) The experimental SAXS data of the Gpc1-dHS (grey dots) and their agreement with the theoretical SAXS profiles calculated for Gpc1 truncated model without the C-terminus (green), the AllosMod best (red), worst (blue), and MES models (orange) respectively. The Gpc1-dC models are shown in (B) & (C); the corresponding Gpc1-dHS models are shown in (E) & (F). The core protein is shown as a light pink cartoon with N-glycans displayed as spheres coloured with the corresponding fit colour in (A). All the models are docked in the ab-initio grey envelope of Gpc1-dC (B & C) and Gpc1-dHS (E & F).

Figure 6 Comparative studies of Gpc1-dC & Gpc1-dHS at different protein concentrations. (A) Analytical SEC profiles showing the concentration effect on the elution volume. One hundred µl of 0.2 mg/ml (light blue), 0.5 mg/ml (blue), 1.4 mg/ml (orange), 3.4 mg/ml (green) & 6 mg/ml (red) solutions of Gpc1-dC and 1mg/ml (violet) & 6 mg/ml (brown) Gpc1-dHS solution were loaded onto the Superdex 200 column. SAXS curves of Gpc1-dC (B) and Gpc1-dHS (C) at protein concentrations of 0.75 mg/ml (orange), 3 mg/ml (green) and 6 mg/ml (red) in a comparison with the SAXS pattern of the equivalent monomer (blue) collected from SEC-SAXS setup. The related P(r) plots are shown in insert boxes in (B) & (C).

Figure 7 Molecular modelling of the Gpc1-dC dimer formed in solution at 6 mg/ml. (A) Cartoon representation of Gpc1-dC (blue) with the glycan chains displayed as red spheres and docked into a transparent grey surface representing the averaged ab-initio model. The fit of the DAMMIN (B) and CORAL (C) models to the Gpc1-dC experimental SAXS data (6mg/ml) is shown, with an insert of the Guinier plot at (B).

Figure 8 SAXS study of deglycosylated Gpc1-dHS: (A) SDS-PAGE gel of Gpc1-dHS and Gpc1-dHS-dN showing a ~5kDa decrease in the size after deglycosylation. (B) Cartoon representation of the enzymatically deglycosylated Gpc1-dHS-dN (green) with the N- & C-terminal displayed as spheres and docked into the transparent averaged ab-initio model. (C) Fit of the Gpc1-dHS-dN dimer model to the experimental SAXS data.

Figure 9 SAXS characterization of non-glycosylated Gpc-1 variants: (A) SDS-PAGE gel of different expressed Gpc-1 mutant with and without disruption of the glycosylation sites before and after the enzymatic removal of HS chains. (B) The SAXS patterns with the DAMMIN fit of Gpc1-WT-dHS (red) and Gpc1-N116Q-dHS (green), with their corresponding Guinier plots as an insert. (C) Intraparticle distance distribution P(r) of Gpc1-WT-dHS (red) and Gpc1-N116Q-dHS (green) with their calculated filtered averaged ab-initio models.

Figure 10 Predicted topology of Gpc-1 on the membrane. (A) Schematic overview of the Gpc1 structure represented in rainbow colours (blue N-terminus, red C-terminus) with N-glycans as red spheres. The model is aligned with a transparent surface showing the sequence conservation of surface residues, coloured as in the inlaid figures. The insert smaller figures show the structural conservation of surface exposed residues in Gpc1- orthologous as generated using ConSurf server (70) and presented by PyMOL as spheres coloured from turquoise (variable) to purple (conserved). (B) The supposed spatial orientation of the Gpc-1 on the cell membrane. The membrane is shown as a grey lipid bilayer with orange GPI anchor connected to the Gpc-1 C-terminus and carried three chains of HS (blue).

TABLES

Table 1 Description of all forms of Gpc1 proteins used in this study

Gpc1 proteins as referred in the text	Characteristics
Gpc1-dC	C-terminally truncated Gpc1 carrying two N-glycans
Gpc1-dC-dN	EndoH deglycosylated C-terminally truncated Gpc1
Gpc1-dHS	Full-length Gpc1 carrying two N-glycans but no HS substitution (by mutagenesis of S486A, S488A, S490A)
Gpc1-dHS-dN	EndoH deglycosylated full-length Gpc1
Gpc1-WT	Wild-type Gpc1 substituted with three HS chains and two N-glycans
Gpc1-WT-dHS	Gpc1-WT treated with HS lyase enzyme
Gpc1-N79Q	Gpc1 substituted with three HS chains and only one glycan at N116 (by mutagenesis of N79Q)
Gpc1-N79Q-dHS	Gpc1-N79Q treated with HS lyase enzyme
Gpc1-N116Q	Gpc1 substituted with three HS chains and only one glycan at N79 (by mutagenesis of N116Q)
Gpc1-N116Q-dHS	Gpc1-N116Q treated with HS lyase enzyme

Table 2 Data collection and refinement statistics

	Dehydrated Gpc1-dHS (4YWT)
Wavelength (Å)	1.000
Resolution range (Å)	43.22 - 2.38 (2.47 - 2.38)
Space group	P 2_1
Unit cell	46.8 166.6 137.7 90 90.4 90
Total reflections	299725 (30301)
Unique reflections	82475 (8185)
Multiplicity	3.6 (3.7)
Completeness (%)	97.9 (97.1)
Mean I/σ(I)	7.44 (1.62)
Wilson B-factor	38.7
R_{merge}(I)	0.108 (0.695)
CC1/2 (I)	0.996 (0.631)
CC* (I)	0.999 (0.880)
R_{work} (F)	0.242 (0.347)
R_{free} (F)	0.278 (0.369)
Number of non-hydrogen atoms	13173
in macromolecules	12823
in ligands	65
water molecules	285
Protein residues	1679
rms deviation from ideal geometry bonds (Å)	0.004
rmsd, angles (^{o})	0.91
Ramachandran favoured (%)	98
Ramachandran outliers (%)	0
Average B-factor	59.1
macromolecules	59.3
ligands	71.1
solvent	46.8

Statistics for the highest-resolution shell are shown in parentheses.

Table 3: SAXS data collection and model parameters of size exclusion experiments

Parameters	Gpc1-dHS	Gpc1-dC
Data Collection parameters		
Beamline	BM29, ESRF	BM29, ESRF
Detector	Pilatus 1M	Pilatus 1M
Beam size at sample (mm^2)	0.7 x 0.7	0.7 x 0.7
Wavelength (Å)	0.9919	0.9919
q range ($Å^{-1}$)	0.1-5	0.1-5
Loaded concentration (mg/ml)	5 mg/ml	6 mg/ml
Structural parameters		
I(0) ($Å^{-1}$), from P(r)	2.4	3.2
R_g(Å), from P(r)	36.6	36.4
I(0) ($Å^{-1}$), from Guinier plot	2.3	3.2
R_g(Å), from Guinier plot	34.2±3.2	35±0.77
D_{max} (Å)	119.5	118.0
Porod volume estimate, V_p ($Å^3$)	95,480	83,810
Average excluded volume, V_{ex} ($Å^3$)	126,100	107,917
Molecular mass (kDa)		
From SAXSMOW	74.9	67.9
Protein sequence	58.5	54
Protein sequence + N-glycans	63.5	59
Modelling parameters		
χ^2 of DAMMIN models	1.14±0.001	0.86±0.004
DAMAVER (20 models) NSD	0.711±0.016	0.644±0.020
AllosMod-best models χ^2	0.95	0.94
Coral χ^2	0.92	0.9
MES χ^2	0.9	0.87
EOM χ^2	0.88	--

Table 4 SAXS data collection and model parameters of different Gpc1 concentration series

Parameters	Gpc1-dC*	Gpc1-dHS**	Gpc1-dHS-dN**	Gpc1-WT-dHS**	Gpc1-N116Q-dHS**
Data Collection parameters					
Beamline	ID 14-3, ESRF	ID 14-3, ESRF	BM29, ESRF	BM29, ESRF	ID 14-3, ESRF
Detector	Pilatus 1M	Pilatus 1M	Pilatus 1M	Pilatus 1M	Pilatus 1M
Beam size at sample (mm^2)	0.7 x 0.7	0.7 x 0.7	0.7 x 0.7	0.7 x 0.7	0.7 x 0.7
Wavelength (Å)	0.931	0.931	0.9919	0.9919	0.931
q range ($Å^{-1}$)	0.1-6	0.1-6	0.1-5	0.1-5	0.1-6
Exposure time (sec)	100 (10x10)	100 (10x10)	20 (10 x 2)	20 (10 x 2)	100 (10x10)
Concentration range (mg/ml)	0.75-6	0.75-6	0.7-2.5	0.7-3	0.7-3
Structural parameters					
I(0) ($Å^{-1}$) , from Guinier plot	10.8±1.5	6.81	10.2	50.5	15.8
R_g(Å), from Guinier plot	60.9±3.5	41.9	54.7	64.5	63.5
I(0) ($Å^{-1}$) , from P(r)	10.7	6.67	10.2	50.4	15.2
R_g(Å), from P(r)	60.7	41.5	55	64.9	63
D_{max} (Å)	210.3	151.7	186	203.8	202.2
Porod volume estimate, V_p ($Å^3$)	169,810	113,380	180,500	333,900	303,580
Excluded volume, V_{ex} ($Å^3$)	195,500	153,390	185,500	346,000	331,900
Molecular mass (kDa)					
From SAXSMoW	110.8	88.3	128.9	168.9	162.3
Monomer from protein sequence	54	58.5	58.5	--	--
Protein sequence + N-glycans	159	63.5	59	--	--
Modelling parameters					
Symmetry	P2	--	P2	P1	P1
$\chi^2$of DAMMIN model	1.4	--	1.06	1.3	1.07
DAMAVER (20 models) NSD	0.89±0.18	--	0.78±0.03	0.97±0.06	0.96±0.06
CORAL χ^2	1.15	--	0.7	--	--

* The processing parameters are for the highest protein concentration sample.

** The processing parameters are for the merged SAXS profile from data collected at different protein concentration.

Table 5. DLS & SAXS data parameters for different Gpc1 concentration series

		R_h (nm) (DLS)	PI (%) (DLS)	MW (kDa) (DLS)	R_g (nm) (SAXS)	MW (kDa) (SAXS)	σ (R_g/R_h)
Gpc1-dC	Monomer	4.06±0.5	11%	89.4±13	3.6	67.9	0.9
	0.75 mg/ml	4.3±0.5	14%	116±12	4.3	70.2	1
	3 mg/ml	4.4±1.1	23%	130±30	5.5	89.1	1.25
	6 mg/ml	5.1±1.1	22%	147±32	6.1	110.8	1.2
Gpc1-dHS	Monomer	4.25±0.6	14%	99.6±27	3.7	74.9	0.87
	0.75 mg/ml	5.2±2	39%	160±62	4.2	86	0.81
	3 mg/ml	5.1±1.9	36.5%	152±55	4.2	86	0.82
	6 mg/ml	5.1±1.7	34%	150±51	4.5	87	0.88

Figures

Figure 1

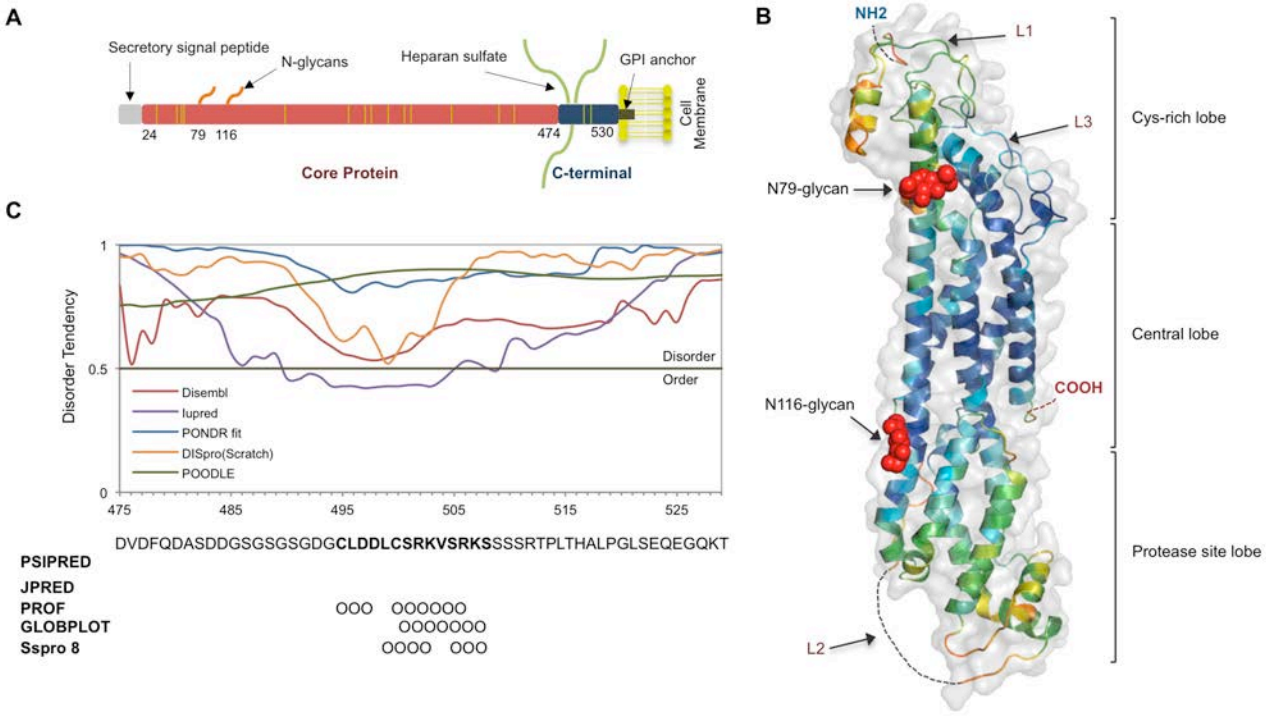


Figure 2

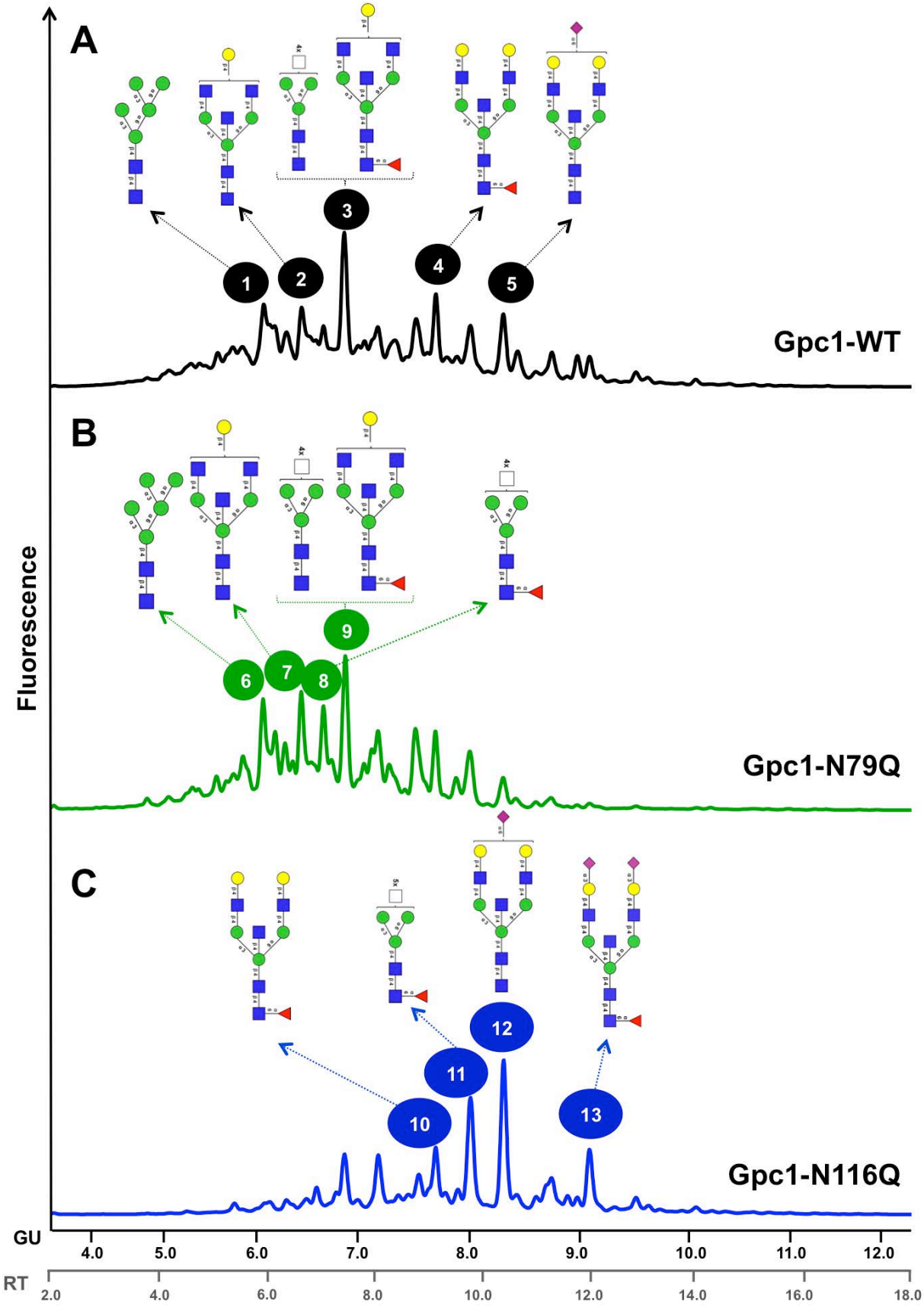


Figure 3

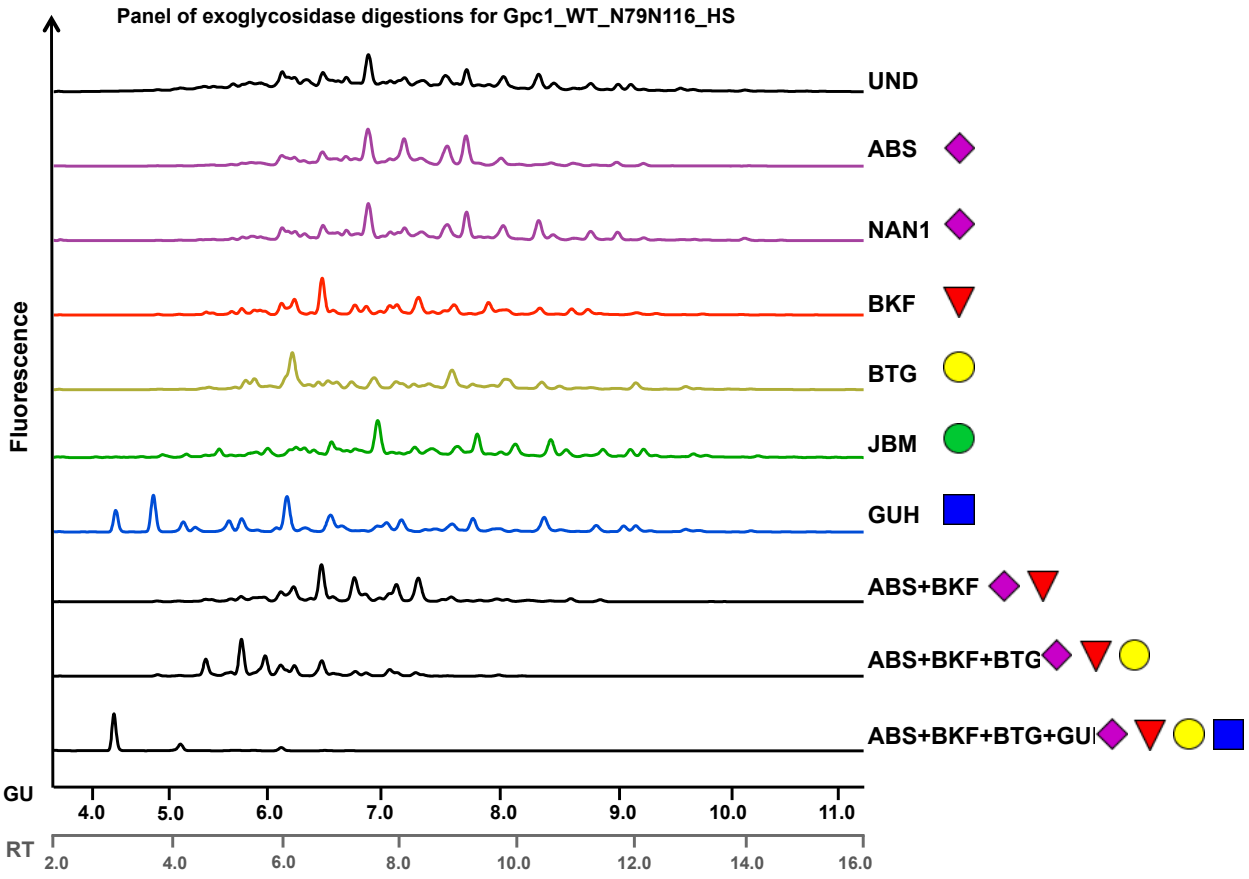


Figure 4

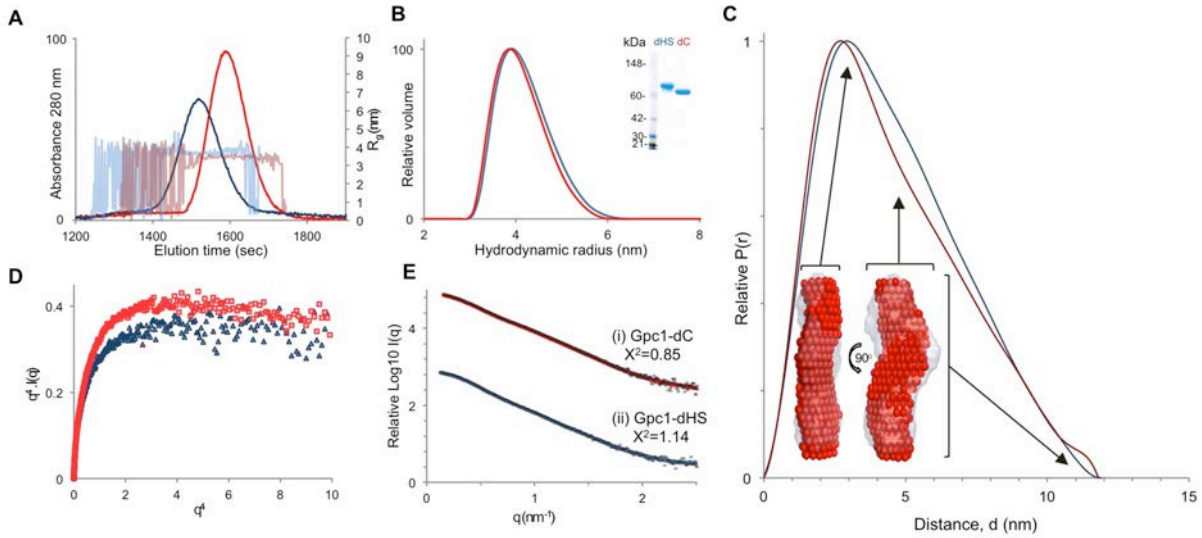


Figure 5

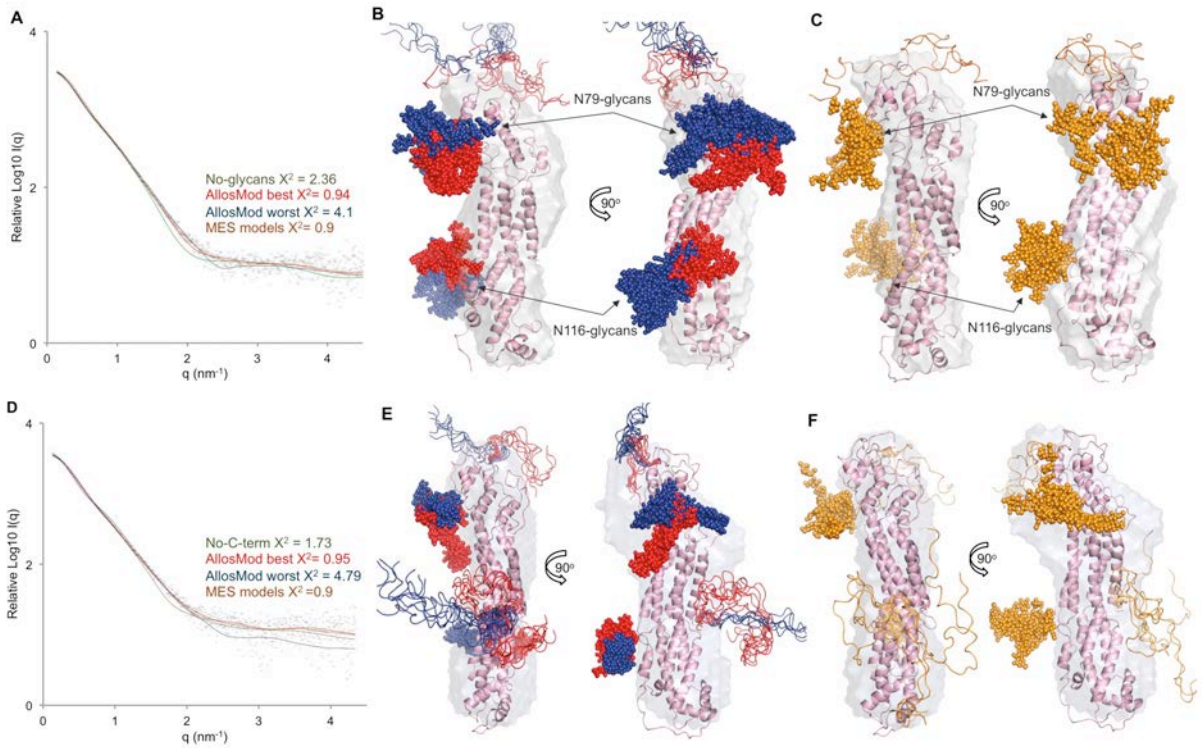


Figure 6

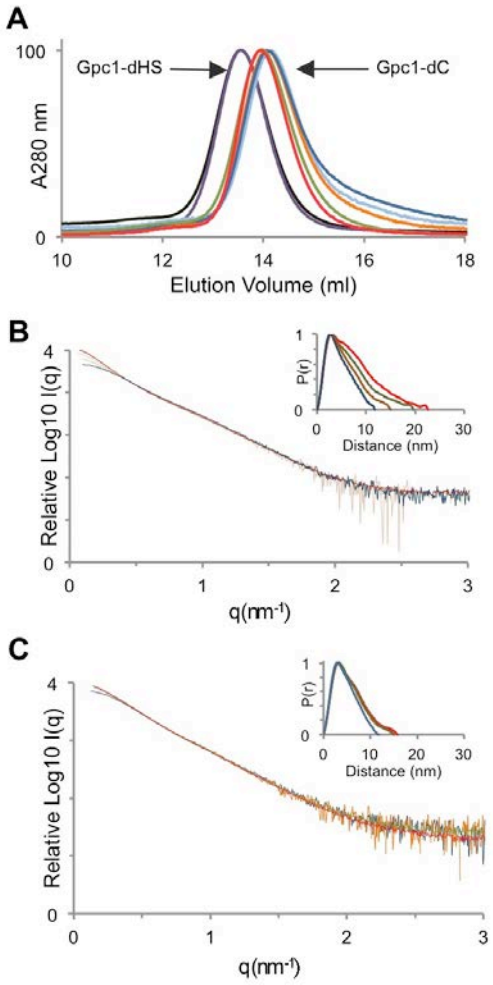


Figure 7

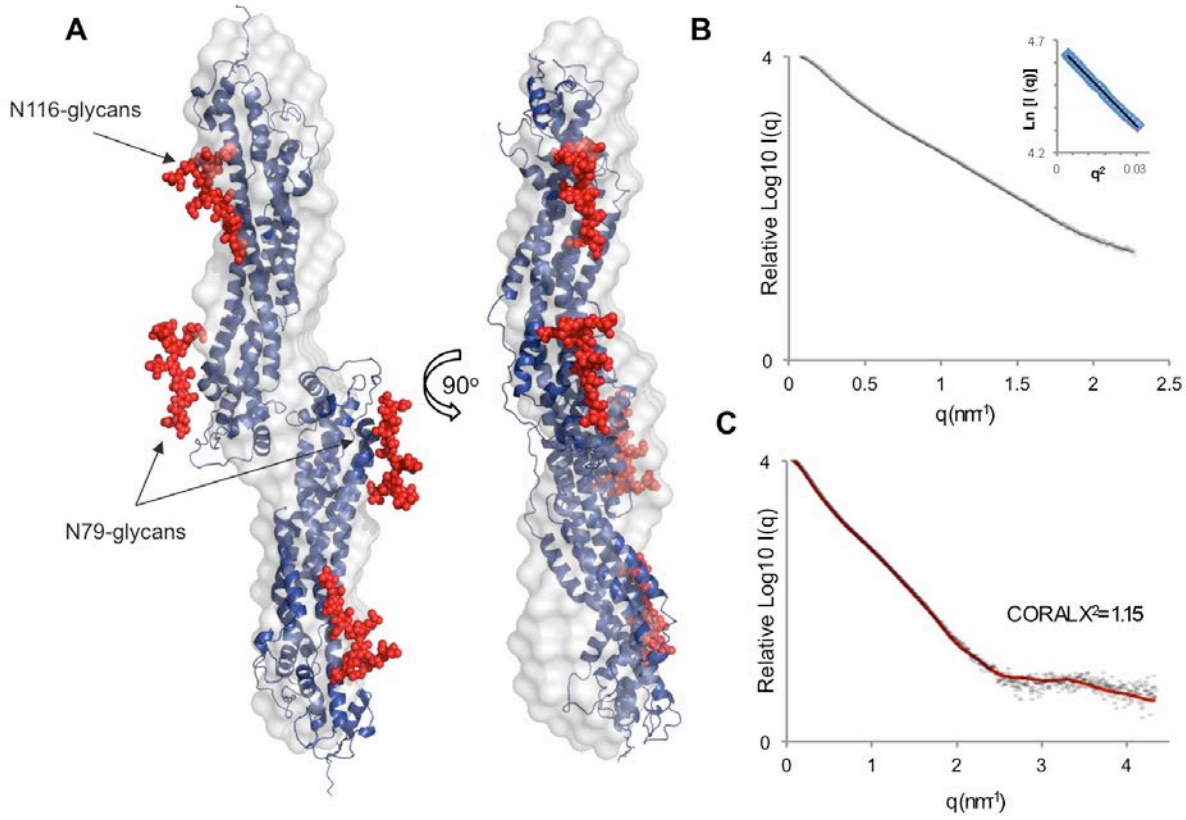


Figure 8

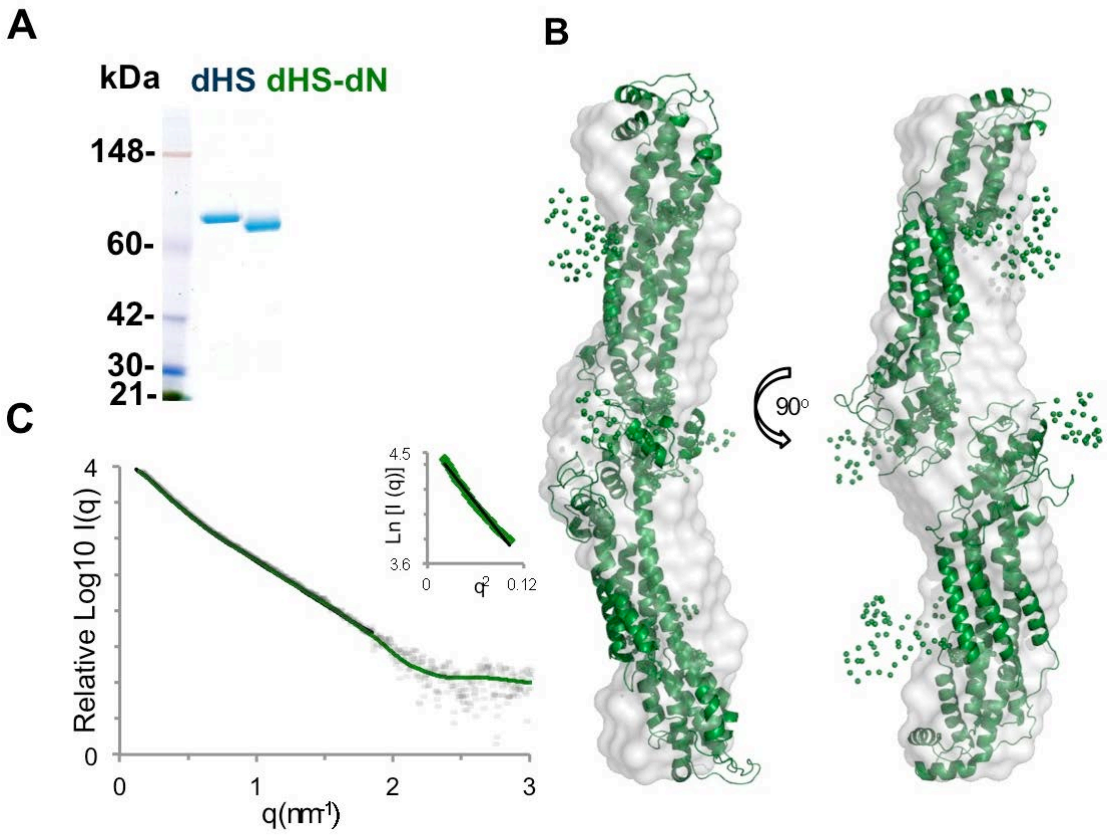


Figure 9

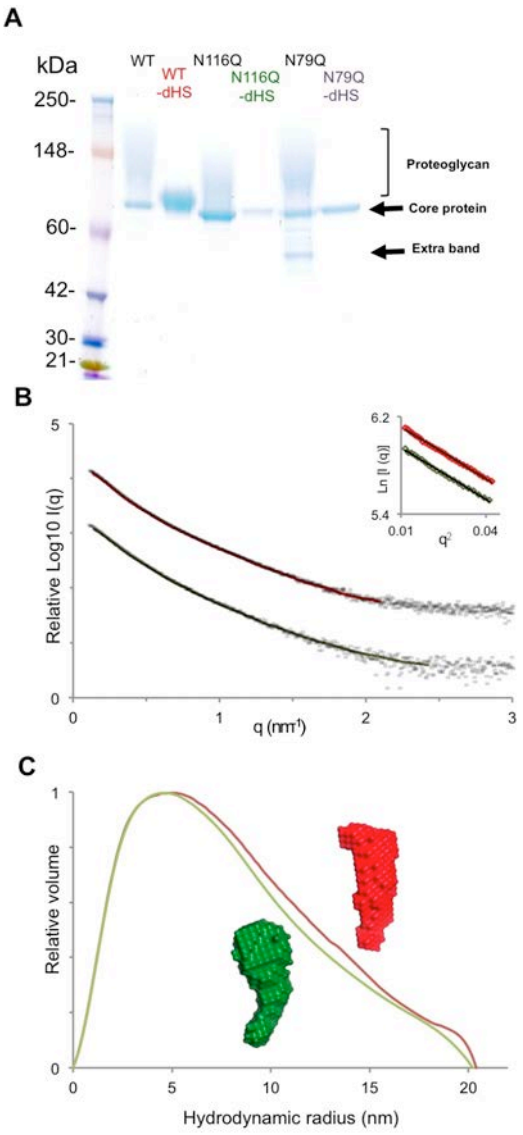
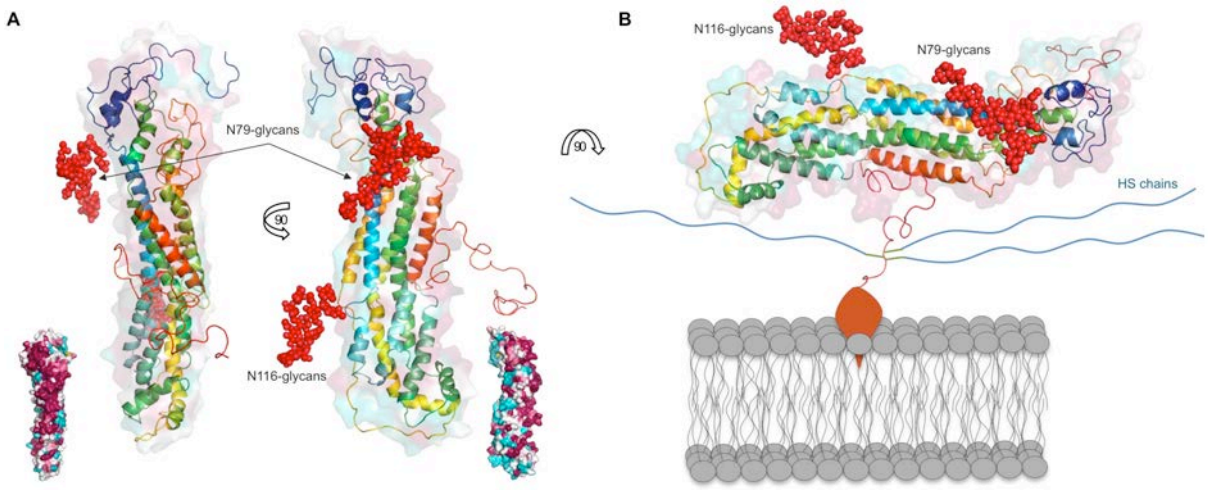
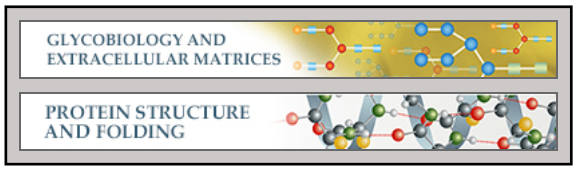


Figure 10



Glycobiology and Extracellular Matrices:

Structural aspects of N-glycosylations and the C-terminal region in human glypican-1



Wael Awad, Barbara Adamczyk, Jessica Örnros, Niclas G. Karlsson, Katrin Mani and Derek T. Logan

J. Biol. Chem. published online July 22, 2015