Evolving views of protein glycosylation

Kurt Drickamer and Maureen E. Taylor

The composition, biosynthesis and known roles of oligosaccharides that are attached to glycoproteins suggest that multiple forces have driven the evolution of proteins that create and recognize these structures. The evolution of glycoprotein biosynthesis and recognition mechanisms can be best understood as a sequential development of functions associated with oligosaccharides.

THE ROLE OF carbohydrates that are attached to proteins and lipids has been a changing one[1,2]. Carbohydrates on the surface of yeast and other free-living eukaryotes probably play largely structural roles, while those on the surface of multicellular eukaryotes can have information-bearing functions in cell–cell recognition and adhesion. Recent work on glycoprotein biosynthesis, which occurs as the molecules make their way to the cell surface, indicates that the carbohydrates also have an important role in intracellular sorting – specifically, in quality control[3,4]. Consideration of the sets of sugars that are involved in these different processes suggests that these functions evolved sequentially and that there were four distinct evolutionary stages, during which different selection pressures selected for distinct types of sugar structure. Here, we summarize evidence from studies of oligosaccharide structure, biosynthesis and recognition that support such a sequential model for the evolution of glycoprotein oligosaccharides (see Fig. 1).

Cores and elaborations in oligosaccharide structures

Given the number of hexoses and the possibilities for assembling them into oligosaccharides by using multiple different linkages, the potential for diversity in oligosaccharides is staggering[5]. However, the number of consituent sugars that are actually present in most groups of oligosaccharides is not large, and the linkages that are used tend to be restricted. In mammals, for instance, projections suggest that synthesis of structures attached to proteins by *N*- and *O*-linkages, together with the glycosidic portions of glycolipids, probably can be accounted for by fewer than 500 glycosyltransferases, each specific for creating a single type of linkage[6,7]. The sugars in mammalian protein-linked oligosaccharides can be divided roughly into two groups: (1) the core sugars (such as *N*-acetylglucosamine, mannose and possibly glucose), which establish the basic branching pattern and are complemented by extensions (usually polylactosamine chains consisting of repeated galactose-*N*-acetylglucosamine disaccharides); and (2) a variety of terminal elaborations, which include galactose, *N*-acetylgalactosamine, L-fucose, sialic acid and sulphates (see Fig. 2).

Biosynthesis

The biosynthetic pathway for construction of N-linked chains on eukaryotic glycoproteins provides evidence for the distinction between core sugars and elaborations[8]. The initial carbohydrate structure that is attached to nascent polypeptides contains two *N*-acetylglucosamine residues, nine mannose residues and three glucose residues, which suggests that glucose should be included in the list of core sugars. This 14-sugar unit is transferred *en bloc* to proteins in the endoplasmic reticulum – a process that seems to be common to all eukaryotic cells. The core is then modified during transit through various luminal compartments; glucose residues, and usually some mannose residues, are removed before the creation of complex structures (Fig. 3).

This rather arcane biosynthetic process might reflect the evolutionary history of the pathway: the *N*-acetylglucosamine-, mannose- and glucose-containing core could be a primordial structure, and galactose, *N*-acetylgalactosamine and other monosaccharides might have been incorporated more recently. Indeed, yeast glycoproteins contain very large high-mannose structures and lack mammalian-type terminal elaborations[9]. Thus, addition of the core appears to reflect this structure's much older role as a cell-wall constituent in lower eukaryotes; the elaborations appear to have evolved

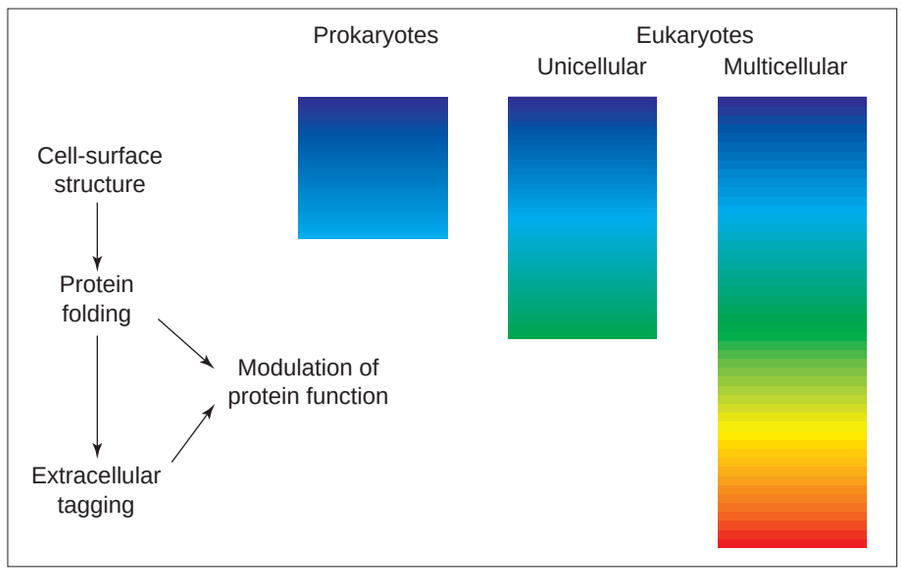

Figure 1
Sequential development of different functions of protein-linked oligosaccharides. The colour scheme represents the changing importance of each function in different groups of organisms. In prokaryotes, glycosylation is associated largely with the cell wall and the outer membrane of Gram-negative bacteria.

K. Drickamer and **M. E. Taylor** are at the Glycobiology Institute, Dept of Biochemistry, University of Oxford, South Parks Rd, Oxford, UK OX1 3QU.
Email: kd@glycob.ox.ac.uk

Copyright © 1998, Elsevier Science Ltd. All rights reserved. 0968–0004/98/$19.00 PII: S0968-0004(98)01246-8

FORMYCON-1082

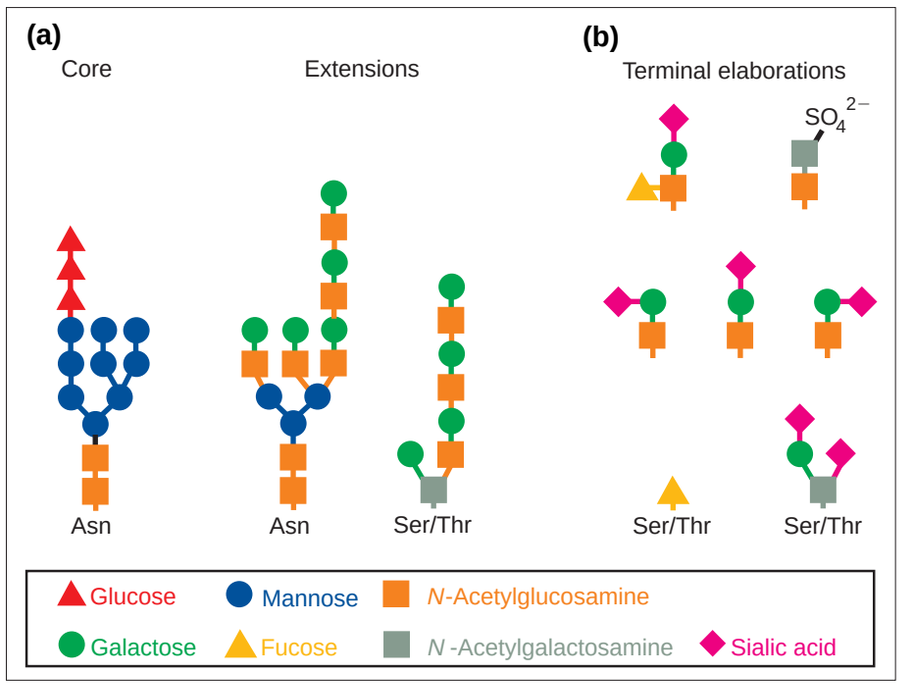

Figure 2
Examples of N- and O-linked glycoprotein oligosaccharides. **(a)** Core sugars and extensions. Extensions include *N*-acetylglucosamine and galactose additions to the core and follow removal of various portions of the core. **(b)** Terminal elaborations. Different terminal elaborations can involve multiple different sugar residues or a common sugar, such as sialic acid, that is linked to the subterminal residue in different ways.

more recently. The terminal variations have expanded extensively within the animal kingdom and have developed in a slightly different way in plants[10].

The intracellular location of the protein glycosylation machinery provides some of the best evidence for the evolving role of the glycosylation pathway. Proteins initially encounter carbohydrate as the glucosylated, high-mannose structure in the endoplasmic reticulum[8]. Sugars associated with the more-recently developed adhesion and recognition functions are added much later in the biosynthetic pathway, particularly in the Golgi. These structures include terminal elaborations of the N-linked glycans, which are added after partial removal of mannose, as well as O-linked sugars. Two aspects of this pathway reflect the evolutionary events that led to its creation. First, the circuitous route of adding a set of nine mannose residues and then removing some of these residues is difficult to explain in conservative energetic terms, but clearly could reflect the grafting of a new pathway onto an old one. Second, the fact that later additions are made in a stepwise manner allows for greater flexibility in the generation of diverse terminal elaborations, while the initial *en bloc* transfer of a common glucosylated, high-mannose oligosaccharide core is more appropriate for creation of more-uniform structures. These differences again reflect the changing evolutionary pressures on the glycosylation pathway.

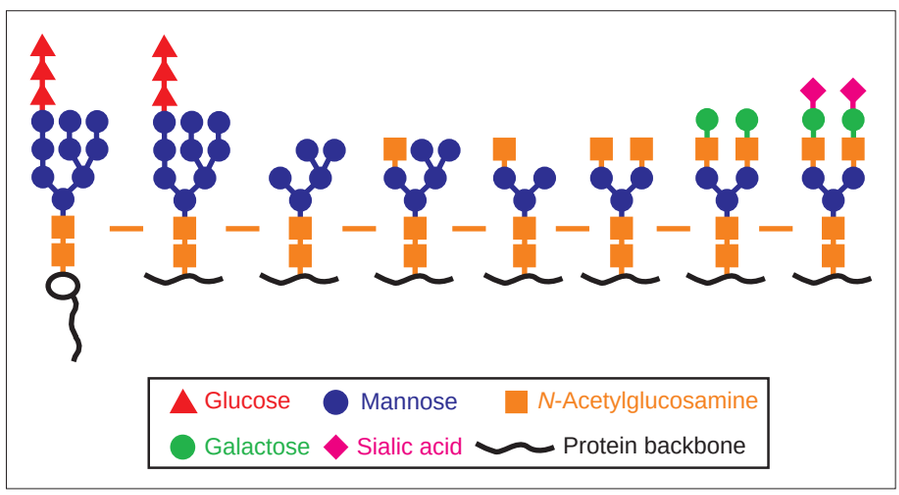

Figure 3
Summary of the biosynthetic pathway for N-linked oligosaccharides. Some of the key steps, starting with *en bloc* transfer of the core from a lipid-linked precursor, are shown. The steps include processing to remove core sugars, which is followed by addition of new terminal elaborations.

Terminal elaborations as recognition markers

Following trimming back of glucose and mannose residues, the remodelling of the core by addition of various types of extensions is largely determined by the addition of *N*-acetylglucosamine residues, which dictate various branching patterns[11]. Extensions often include addition of galactose, either as a single residue at the end of a branch or as a longer polylactosamine chain. These extended structures then serve as a scaffold for addition of a variety of terminal elaborations. These elaborations can be simple: sialic acid, for example, can be attached through various different linkages. Alternatively, the elaborations can be more complex: additional residues can be added sequentially to form either linear (e.g. attachment of *N*-acetylgalactosamine followed by attachment of a sulphate group) or branched arrangements (e.g. attachment of fucose and sialic acid). The examples shown in Fig. 2 represent only a small subset of the multitude of known elaborations, many of which were originally recognized as unique blood-group-specific antigens.

Comparison of the structures and biosynthesis of N-linked sugars with those of O-linked structures suggests that O-linked sugars can be viewed largely as extensions and terminal elaborations. The O-linked sugars can simply be single sugar residues. Alternatively, complex structures can be built on an initial fucose[12], *N*-acetylgalactosamine[13] or mannose residue[14] that is attached to the polypeptide. A common structure is a polylactosamine scaffold similar to that sometimes found on N-linked sugars. As in the case of the extensions and terminal elaborations added to N-linked sugars, additions are single and stepwise, and occur relatively late in the biosynthetic pathway, in the Golgi rather

than in the endoplasmic reticulum. There is no *en bloc* transfer of a common core onto serine/threonine residues; this is consistent with the view that the O-linked structures are analogous to the extensions and elaborations of the N-linked oligosaccharides. These considerations suggest that the O-linked structures on mammalian glycoproteins developed in parallel with the more recent part of the N-linked glycosylation pathway.

While the roles of many of the terminal structures on glycoprotein oligosaccharides are not well understood, selective recognition of some of these sugars by endogenous receptors suggests that these structures serve as readily accessible tags on the surface of glycoproteins. One of the simplest modifications to the N-linked core is the addition of a phosphate group, which targets hydrolases to lysosomes[15]. This post-Golgi, but intracellular, sorting function, which is mediated by mannose-6-phosphate receptors, is an example of the use of unique terminal sugars in sorting. The existence of this sorting function in metazoans but not in yeast and dictyostelium[16] indicates that the function is a more-recent evolutionary development (Fig. 1).

Many additional sugar-specific receptors recognize glycoprotein oligosaccharides once these sugars reach the cell surface[1]. Two main roles of these lectins are in clearance of proteins from the circulation and in cell–cell adhesion. Clearance of asialoglycoproteins that bear terminal galactose residues and clearance of glycoprotein hormones that bear terminal *N*-acetylgalactosamine-4-sulphate residues are two of the most-extensively studied extracellular sorting events[17]. Two families of sugar-specific adhesion receptors, the selectins and the sialoadhesins, interact with sialylated cell-surface ligands[18,19]. While these recognition functions are becoming increasingly well understood at the molecular level, the diversity of terminal oligosaccharide structures found on different proteins suggests that there are many as-yet-unidentified receptors. These recognition functions would be useful only in multicellular organisms and thus appeared relatively recently in evolution (Fig. 1). Another group of extracellular mammalian lectins has also evolved: these recognize core sugars on bacteria and fungi, and target an innate immune response towards these organisms that is based on the preponderance of core sugars compared to the lower abundance and internal position of core sugars at the surface of mammalian cells[1].

An intermediate role for sugars

The early attachment of core N-linked structures to glycoproteins suggests that this part of the sugar functions early in the life of the glycoconjugate. This argument is particularly appealing in the case of the glucose component, because these residues are transient and are lost once the glycoprotein reaches the later portion of the Golgi. Recent evidence suggests that removal of glucose from the core structure indicates to the sorting apparatus in the luminal compartments that the protein component is appropriately folded[3,4]. These sorting steps are mediated by two intracellular lectins: the chaperones calnexin and calreticulin. The presence of at least some components of this sorting pathway in relatively simple eukaryotes, such as yeast[20], but not in prokaryotes suggests that the sorting role of sugars evolved relatively early but probably after the structural role (Fig. 1). While it is possible to argue that the sorting function of sugars preceded their structural role, there is no evidence that yeast calnexin monitors protein folding. The role of glycosylation in protein folding must therefore have evolved after the structural role was established

Additional intracellular lectins might interact with the core sugars during the early phases of transit through the luminal compartment. Protein ERGIC-53, which shuttles between the endoplasmic reticulum and the early compartments of the Golgi, is one of a family of intracellular animal proteins that resemble legume lectins and appears to bind mannose-containing structures[21]. The fact that the structures recognized by both this L-type-lectin family and the calenexin–calreticulin group are of the core type (glucose and mannose) is consistent with the intracellular sorting function of sugars evolving after their original structural role but before their role in cell-surface interactions.

Collateral functions of oligosaccharides

Studies of the effects of glycosylation on protein structure and function suggest that glycosylation can affect the behaviour of proteins in a cell-free context. At the structural level, attachment of carbohydrate can decrease amino acid backbone mobility throughout the protein and increase thermal stability[22]. In some instances, structural analysis has revealed at least part of the underlying physical basis for these effects, in the form of specific contacts with the sugar residue[23,24].

The structures of oligosaccharides that are attached to glycoproteins provide interesting insight into the potential importance of carbohydrate in glycoprotein structure and function. Comparison of the three-dimensional structures of two of the many distinct trypanosome variant surface glycoproteins that can coat the parasite reveals that, in one case, the hydrophobic core near the base of the protein is covered by a short stretch of peptide that adopts an α-helical conformation; in another case, this patch is instead covered by a portion of a high-mannose oligosaccharide that is attached to a nearby asparagine residue[25].

This comparison suggests an important principle of glycoprotein evolution: if a glycosylation-directing sequence that is appropriately accessible on the surface of a protein appears, an oligosaccharide will be attached by the pre-existing glycosylation machinery; the selection process will then work on the glycosylated protein. In the case of the two trypanosome variant surface glycoproteins, if the oligosaccharide is present, the α-helical segment can be deleted without disrupting the protein structure. At this point, the glycosylation becomes an essential part of the structure of this form of the protein. This fact leads to the hypothesis that, once the glycosylation machinery was in place, attached oligosaccharides inevitably came to play essential roles in the functions of some glycoproteins. The portions of N-linked oligosaccharides that are involved in such functions seem to be core residues. It often does not matter which particular sugars are attached to a protein; the protein just has to be glycosylated on a particular asparagine residue[22]. This phenomenon is probably due to the proximity of the core to the protein surface. By contrast, terminal elaborations are distally located, which facilitates their interaction with receptors.

Viewed in this way, direct effects of glycosylation on the behaviour of glycoproteins appear to be a relatively recently evolved function of sugars (Fig. 1). This view helps to explain the otherwise-baffling fact that sugar residues sometimes perform functions that apparently could be accomplished by amino acid residues. It has been difficult to understand how such functions could have driven the evolution of the complex glycosylation machinery. It is far more likely that evolution of the machinery was driven by the evolving functions discussed above, while the structural effects

of glycosylation arose adventitiously and were subsequently selected for.

Conclusions and outlook

The sequential development of different glycoprotein functions we propose is necessarily somewhat speculative and not readily subject to direct experimental verification. Nevertheless, this scheme provides a reasonably satisfying explanation for a number of seemingly anomalous aspects of glycoprotein biosynthesis and a sometimes-bewildering multiplicity of proposed functions. As we learn more about the diversity of glycoprotein structure and the genetic complexity that underlies the biosynthetic machinery, we should be able to discern the evolving role of complex carbohydrates more clearly.

Acknowledgements

We thank Roger Dodd for help with preparation of figures and the Wellcome Trust and the Oxford Glycobiology Endowment for funding.

References

1 Drickamer, K. and Taylor, M. E. (1993) *Annu. Rev. Cell Biol.* 9, 237–264
2 Gahmberg, C. G. and Tolvanen, M. (1996) *Trends Biochem. Sci.* 21, 308–311
3 Hebert, D. N., Simons, J. F., Peterson, J. R. and Helenius, A. (1995) *Cold Spring Harbor Symp. Quant. Biol.* 60, 405–415
4 Fiedler, K. and Simons, K. (1995) *Cell* 81, 309–312
5 Hughes, R. C. (1983) *Glycoproteins*, Chapman and Hall
6 Kleene, R. and Berger, E. G. (1993) *Biochim. Biophys. Acta* 1154, 283–325
7 Natsuka, S. and Lowe, J. B. (1994) *Curr. Opin. Struct. Biol.* 4, 683–691
8 Kornfeld, R. and Kornfeld, S. (1985) *Annu. Rev. Biochem.* 54, 631–664
9 Kukuruzinska, M. A., Bergh, M. L. E. and Jackson, B. J. (1987) *Annu. Rev. Biochem.* 56, 915–944
10 Driouich, A., Faye, L. and Staehelin, L. A. (1993) *Trends Biochem. Sci.* 18, 210–214
11 Schachter, H. (1991) *Glycobiology* 1, 453–461
12 Moloney, D. J., Lin, A. I. and Haltiwanger, R. S. (1997) *J. Biol. Chem.* 272, 19046–19050
13 Marth, J. D. (1996) *Glycobiology* 6, 701–705
14 Yuen, C-T. *et al.* (1997) *J. Biol. Chem.* 272, 8924–8931
15 Kornfeld, S. (1992) *Annu. Rev. Biochem.* 61, 307–330
16 Mehta, D. P. *et al.* (1996) *J. Biol. Chem.* 271, 10897–10903
17 Drickamer, K. (1991) *Cell* 67, 1029–1032
18 Lasky, L. A. (1995) *Annu. Rev. Biochem.* 64, 113–139
19 Powell, L. D. and Varki, A. (1995) *J. Biol. Chem.* 270, 14243–14246
20 Parlati, F., Dominguez, M., Bergeron, J. J. M. and Thomas, D. Y. (1995) *J. Biol. Chem.* 270, 244–253
21 Itin, C., Roche, A. C., Monsigny, M. and Hauri, H. P. (1996) *Mol. Biol. Cell* 7, 483–493
22 Dwek, R. A. (1995) *Biochem. Soc. Trans.* 23, 1–25
23 Mer, G., Hietter, H. and Lefèvre, J-F. (1996) *Nat. Struct. Biol.* 3, 45–53
24 Thijssen-van Zuylen, C. W. E. M. *et al.* (1998) *Biochemistry* 37, 1933–1940
25 Blum, M. L. *et al.* (1993) *Nature* 362, 603–609

Conservation of gene order: a fingerprint of proteins that physically interact

Thomas Dandekar, Berend Snel, Martijn Huynen and Peer Bork

A systematic comparison of nine bacterial and archaeal genomes reveals a low level of gene-order (and operon architecture) conservation. Nevertheless, a number of gene pairs are conserved. The proteins encoded by conserved gene pairs appear to interact physically. This observation can therefore be used to predict functions of, and interactions between, prokaryotic gene products.

COMPLETELY SEQUENCED GENOMES provide us with an opportunity to study the evolution of genome organization at a comprehensive level. A variety of studies have focused on the conservation of gene order in evolution, and the authors have drawn different conclusions, depending on the phylogenetic distance between the species compared and on the genes that were analyzed[1–5]. For example, conservation of gene order between *Mycoplasma genitalium* and *Mycoplasma pneumoniae*[6] is likely to be a result of a lack of time for genome rearrangements after divergence of the two organisms from their last common ancestor. Hence, if one is interested in the selective constraints that preserve gene order, only relatively long evolutionary distances between the species compared should be considered. However, the distances should be small enough that a significant number of orthologous genes is still shared by the species.

Gene order is already considerably disrupted when the protein-sequence identity shared by orthologs in two genomes is <50%[7]. We therefore analyzed genes from three sets of three completely sequenced genomes for which at least two of the intergenomic distances show less than 50% identity in shared orthologs (Fig. 1), which should be a sufficient test set for systematic studies. The genome sequences used (see Box 1) included those of proteobacteria (*Escherichia coli*[8], *Haemophilus influenzae*[9] and *Helicobacter pylori*[10]), Gram-positive bacteria (*M. genitalium*[11], *M. pneumoniae*[12] and *Bacillus subtilis*[13]) and archaea (*Methanococcus jannaschii*[14], *Methanobacterium thermoautotrophicum*[15] and *Archaeoglobus fulgidus*[16]).

To ensure that any conservation of gene order dates back to the earliest point at which the sequences compared diverged (rather than to more recent horizontal gene-transfer events) and hence reflects evolutionary constraints, we only considered genes that show the same order in a set of three genomes. For example, the urease operon is present in both *H. influenzae* and *H. pylori*, but is absent from *E. coli*. In *H. influenzae*, the G–C content of the urease-operon

T. Dandekar, **B. Snel**, **M. Huynen** and **P. Bork** are at the European Molecular Biology Laboratory, Postfach 102209, D-69012 Heidelberg, Germany; and **T. Dandekar**, **M. Huynen** and **P. Bork** are also at the Max-Delbrück-Centrum fuer Molekulare Medizin, Robert-Roessle Str. 10, 13122 Berlin-Buch, Germany.
Email: huynen@embl-heidelberg.de

Copyright © 1998, Elsevier Science Ltd. All rights reserved. 0968–0004/98/$19.00 PII: S0968-0004(98)01274-2