# EXHIBIT Y
# PART 1

Eur. J. Biochem. *218*, 1–27 (1993)


*Review*

# Protein glycosylation
## Structural and functional aspects

Halina LIS and Nathan SHARON

Department of Membrane Research and Biophysics, The Weizmann Institute of Science, Rehovot, Israel

(Received April 16/July 15, 1993) – EJB 93 0558/0

During the last decade, there have been enormous advances in our knowledge of glycoproteins and the stage has been set for the biotechnological production of many of them for therapeutic use. These advances are reviewed, with special emphasis on the structure and function of the glycoproteins (excluding the proteoglycans). Current methods for structural analysis of glycoproteins are surveyed, as are novel carbohydrate–peptide linking groups, and mono- and oligo-saccharide constituents found in these macromolecules. The possible roles of the carbohydrate units in modulating the physicochemical and biological properties of the parent proteins are discussed, and evidence is presented on their roles as recognition determinants between molecules and cells, or cell and cells. Finally, examples are given of changes that occur in the carbohydrates of soluble and cell-surface glycoproteins during differentiation, growth and malignancy, which further highlight the important role of these substances in health and disease.

Among the different types of covalent modifications that newly synthesized proteins undergo in living organisms, none is as common as glycosylation [1–6]. It is also the most diverse, both with respect to the kinds of amino acid that are modified and the structures attached. The origins for this diversity are chemical as well as biological. The former results from the ability of monosaccharides to combine with each other in a variety of ways that differ not only in sequence and chain length, but also in anomery ($\alpha$ or $\beta$), position of linkages and branching points. Further structural diversification may occur by covalent attachment of sulfate, phosphate, acetyl or methyl groups to the sugars. Therefore, in theory, an enormous variety of glycans, both oligosaccharides and polysaccharides, can be generated from a relatively limited number of monosaccharides. Biological diversity derives from the fact that, whereas proteins are primary gene products, glycans are secondary gene products. As a result, glycosylation is species- and cell-specific, and is determined as well by the structure of the protein backbone and the carbohydrate attachment site. This means that glycosylation of any protein is dependent on the cell or tissue in which it is produced and that the polypeptide encodes information that directs its own pattern of glycosylation.

In an individual glycoprotein more than one carbohydrate unit is often present, attached at different positions by either an N-linkage, an O-linkage or both. Moreover, each attachment site frequently accommodates different glycans, a phenomenon referred to as site heterogeneity. This results in microheterogeneity of the whole molecule and creates discrete subsets, or glycoforms, of a glycoprotein that have different physical and biochemical properties, which, *in turn*, may lead to functional diversity [7]. In short, glycosylation of a polypeptide usually generates a set of glycoforms, all of which share an identical backbone but are dissimilar either in the structure or disposition of their carbohydrate units or in both. The earlier view that heterogeneity of the carbohydrates of glycoproteins is random, mainly due to the lack of fidelity in their synthesis, seems no longer tenable, since the relative proportions of such glycoforms appear to be reproducible and highly regulated. They depend on the cellular environment in which the protein is glycosylated and may therefore vary with the type, as well as the physiological state, of the organism, tissue or cell in which the glycoprotein is made.

The ubiquity of glycosylation is well established. It occurs without exception in integral membrane proteins of higher organisms and is quite common with secretory proteins. For instance, in blood serum, almost all proteins are glycosylated, as are those in hen egg white. Glycoproteins are now known to occur also in the cytoplasm and nucleus [8]. Whereas bacteria were for a long time considered to lack the ability to synthesize glycoproteins, this now appears not to be the case. Many species of archaebacteria, as well as of eubacteria, produce glycoproteins, although mostly of types not found in other organisms [9–11].

During the last decade, there has been a vast expansion in our knowledge of the distribution of glycoproteins in nature, and of their structure, biochemistry and biosynthesis; important insights have also been obtained into their roles. Detection and isolation of glycoproteins have been facili-

*Correspondence to* N. Sharon, Department of Membrane Research and Biophysics, The Weizmann Institute of Science, IL-76100 Rehovot, Israel

*Abbreviations.* CHO, Chinese hamster ovary; GlPtdIns, glycosyl-phosphatidylinositol; GM-CSF granulocyte-macrophage colony stimulating factor; hCG, human chorionic gonadotropin; LH, lutotropin; N-CAM, neural cell adhesion molecule; tPA, tissue plasminogen activator.

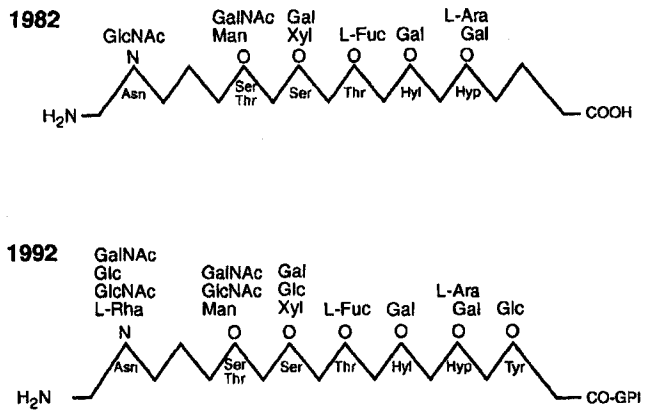



**Fig. 1. Protein–carbohydrate linkages known in 1982 and in 1992.** GPI stands for glycosyl-phosphatidylinositol.

tated, not the least thanks to the availability of an increasing range of lectins with a wide spectrum of specificities which are capable of distinguishing subtle differences in the structure of oligosaccharide units of glycoproteins [12–14]. Refinements of known separation and analytical methods, as well as introduction of new technologies, have made it possible to determine complex glycan structures at the nanomole, and sometimes even picomole, level in relatively short periods of time. As a consequence, the number of known structures of carbohydrate units of glycoproteins has grown immensely and the early assumption, that living organisms form only an exceedingly small fraction of the theoretically possible molecular permutations of the dozen or so monosaccharides typically found in glycoconjugates, seems no longer justified. Not only have novel structures been discovered, but so too have new monosaccharide constituents and new linkages between the peptide backbone and the carbohydrate unit. The latter point is illustrated in Fig. 1 which compares the linkages known today with those known a decade earlier. The novel linkages include, in addition to hitherto unknown N- and O-glycosidic bonds, the glycosyl-phosphatidylinositol (GlPtdIns) anchor, a new class of widely occurring linkage, where the carbohydrate is attached to the C-terminal amino acid of the protein via ethanolamine phosphate [15–18]. It should be noted, however, that this kind of attachment of carbohydrate to the protein is not a glycosylation process in the strict sense, since the sugar is not bound to the polypeptide chain by a glycosidic linkage; it has been termed 'glypiation'.

Striking advances have been made in synthetic carbohydrate chemistry. Linear or branched oligosaccharides consisting of up to a dozen units, as well as different glycopeptides, can now be produced in the laboratory; these include constituents of N- and O-glycoproteins and of the GlPtdIns anchor [19, 20]. Simple procedures for enzymatic synthesis of oligosaccharides, at a hundred milligram scale using immobilized enzymes, have also become available [21, 22]. Nevertheless, synthesis of most oligosaccharides found in glycoproteins is still difficult (or impossible), as is the scaling up of the synthetic procedures to the gram level. The synthetic products are widely employed as reference compounds, for the investigation of specificity and structure/function relationships of enzymes, lectins, antibodies, etc. Their application as potential drugs, e.g. for prevention of microbial infections or inflammation, is under intensive investigation. Conjugation of oligosaccharides of known structure to proteins (e.g. bovine serum albumin) affords 'neoglycoproteins' with desirable carbohydrate units [23]. These compounds too are useful for probing the specificity of carbohydrate-binding proteins and as affinity matrices for the isolation of such proteins. In addition, they serve as immunogens for the production of antibodies against oligosaccharides and in studies of the role of the carbohydrate in glycoproteins.

Progress has been made in our knowledge of the three-dimensional structures of oligosaccharides, both free and protein-linked, based on nuclear magnetic resonance (NMR), various modelling techniques and X-ray crystallography [24–30]. It has become apparent that, in solution, the oligosaccharides are flexible molecules that can adopt different conformations, only a few of which are recognized by carbohydrate-specific proteins.

The principal biosynthetic pathways leading to the production of mature glycoproteins by glycosyltransferases, glycosidases and carbohydrate-modifying enzymes, and in particular the fine details of the dolichol phosphate cycle, in which the $Glc_3Man_9(GlcNAc)_2$ precursor of the commonly occurring N-oligosaccharides is synthesized, have been known for some time [31]. While much attention is still being given to purification and characterization of the enzymes involved, and to the reactions they catalyze, emphasis has shifted to topological aspects, control mechanisms and molecular biology of glycosylation. Under intense investigation are problems such as subcellular sites of glycosylation, translocation of sugars from the cytoplasmic face to the lumen of the endoplasmic reticulum and the Golgi apparatus, trafficking between organelles (e.g. from the Golgi to lysosomes) and, most importantly, regulation of glycoprotein processing and maturation. Much of our knowledge in these areas has been obtained with the aid of mutant mammalian cell lines, selected mostly by virtue of their resistance to the toxic action of lectins and shown to be deficient in certain enzymes involved in individual steps of protein glycosylation [32]. Another source of information comes from the use of specific inhibitors of transferases (e.g. tunicamycin) and of glycosidases (such as castanospermine, nojirimycin and swainsonine) [33, 34]. Mapping of the subcellar sites of protein glycosylation is aided by the use of lectins and of antibodies to purified glycosylated enzymes [35, 36].

New approaches became available with the emergence of genetic engineering techniques. For instance, oligonucleotide-directed mutagenesis allows for specific changes in the primary structure of glycoproteins and facilitates the examination of factors governing site-specific glycosylation and oligosaccharide processing. In glycoproteins with more than one glycan, mutagenesis provides insights into the contribution of each glycan to the overall properties of the molecule.

Evidence has accumulated for the existence of proteins which mediate the transport of sugar nucleotides across the membranes of the endoplasmic reticulum and the Golgi apparatus [37, 38]. The transporters, or antiporters, facilitate the entry of the sugar donor into the lumen of these organelles in a reaction coupled to the equimolar exit of the corresponding nucleoside monophosphate. Some of the transporters have been partially purified and shown to be both organelle- and substrate-specific.

The role of the Golgi complex in the ordered remodelling of N-oligosaccharide chains and the biosynthesis of O-glycans is firmly established [35, 37]. It has also become clear that this organelle consists of a series of functionally distinct compartments: *cis*, medial and *trans*. As glycoproteins pass

through these compartments, they acquire their O-units, while the N-oligosaccharides, the precursor of which is attached to the growing polypeptide chain in the endoplasmic reticulum, undergo a series of sequential trimming and elongation reactions, as if on an assembly line [36]. The genes coding for the relevant enzymes are being cloned and sequenced at an increasingly fast rate; studies on their regulation are in progress [39–42]. A recent milestone is the cloning and sequencing of the cDNAs which code for the glycosyltransferases that determine human blood types $A_1$ and B ($\alpha$-1,3-*N*-acetylgalactosaminyltransferase and $\alpha$-1,3-galactosyltransferase, respectively), and of the corresponding cDNA from cells of the H(O) type [43]. The deduced sequences for the two transferases differ only in four amino acid residues, while a critical single base deletion found in the O individuals is predicted to give rise to an entirely different protein which would be expected to be nonfunctional.

Several glycosyltransferases have been shown to exhibit branch specificity, which accounts for the marked differences in chain length sometimes found between different branches of the same glycan [44]. An insight into the mechanism by which the protein backbone may control glycosylation was provided by the finding that, in addition to the combining site(s) for the sugar donor and acceptor, glycosyltransferases can contain a site that recognizes certain features in the peptide moiety of the acceptor glycoprotein [45].

Recent work has clarified several aspects of the catabolism of N-glycoproteins [46]. In this process, a series of lysosomal enzymes act in a highly ordered manner to ensure the complete degradation of glycoproteins. It is achieved by stepwise hydrolysis of the major portion of their glycans by a set of exo-glycosidases, followed by the disassembly of the protein and the carbohydrate–peptide linkage region. An alternative pathway for the degradation of glycoproteins starts with proteolysis of the polypeptide backbone and involves the participation of specific endoglycosidase(s) [47, 48]. The physiological importance of high precision in the lysosomal degradative system is clearly illustrated by the occurrence of serious, often fatal, disorders in individuals with genetic defects in glycosidase production [49, 50]. Details of the degradation of O-glycans are largely unknown but, as recently shown, in this case, too, genetic defects in glycosidase production may lead to serious disorders [51].

The fact that the carbohydrate units of glycoproteins have been conserved in evolution and the growing awareness of the widespread occurrence and structural diversity of glycoproteins, coupled with the realization that oligosaccharide structures of glycoproteins sometimes undergo dramatic changes with differentiation and in pathological processes, intensified the search for their biological role(s). The ability of the carbohydrate groups to modulate the physical properties of the protein to which they are attached, especially the overall folding of the nascent polypeptide chain, as well as to protect it against proteolysis, is well documented [1–4]. More importantly, there is increasing evidence for the concept, formulated over 20 years ago [52], that carbohydrates act as recognition determinants in a variety of physiological and pathological processes [7, 53–57]. These include clearance of glycoproteins from the circulatory system [52], intracellular trafficking of enzymes [58] and a wide range of cell–cell interactions, from the attachment of sperm to ova [59], to adhesion of infectious microorganisms to host tissues [60–62]. Particularly exciting is the recent demonstration that binding of carbohydrates on the surface of leukocytes, with a class of animal lectins designated 'selectins', controls leukocyte traffic by mediating adhesion of these cells to restricted portions of the endothelium and their recruitment to inflammatory sites [63–65]. Within a short period of time, the study of selectins and their receptors has become, perhaps, the most active area in glycobiology. Intensive attempts are in progress to design carbohydrate-based selectin inhibitors, which, in turn, may be candidates for a new class of anti-inflammatory drugs [66].

**Table 1. Some glycosylated proteins of therapeutic interest.** This table is based largely on the review of Rasmussen [67].

| Glycoprotein | Carbohydrate-peptide linkage | Requirement of carbohydrate for activity | Biotechnology production | Clinical use |
|---|---|---|---|---|
| $\alpha_1$-Antitrypsin | N | | + | |
| Coagulation factor VIII | N and O | | ± | + |
| Erythropoietin | N and O | ± | + | + |
| Follicle stimulating hormone | N and O | + | | |
| Glucocerebrosidase | N | | + | + |
| Granulocyte colony-stimulating factor | O | | + | + |
| Granulocyte-macrophage colony-stimulating factor | N and O | – | + | + |
| Human chorionic gonadotropin | N and O | + | | |
| Interleukin-2 | O | – | | |
| Interferon-$\beta$ | N | – | + | + |
| Interferon-$\gamma$ | N | – | + | + |
| Protein C | N | | | |
| Soluble CD4 | N | – | + | |
| Tissue plasminogen activator | N | ± | + | + |

This is one example of how increased knowledge of carbohydrate structure and function might be utilized for therapeutic purposes. Another example is the enzymatic modification of the glycan of the enzyme glucocerebrosidase (glucosylceramidase), which is essential for its clinical use (under the trade name Ceredase) in the treatment of patients with Gaucher's disease [68, 69]. It is the first, and thus far probably the only, case of enzyme replacement therapy, a concept suggested some 30 years ago. Also, genetic engineering makes it possible to produce glycoproteins in heterologous systems on a large scale, both for research purposes and for therapeutic use (Table 1). We are indeed witnessing the emergence of glycotechnology [70], a branch of biotechnology that uses novel approaches to manipulate carbohydrates or related materials, with the aim of creating new products or new procedures for the betterment of our lives. An impressive example is erythropoietin, a circulating glycoprotein hormone that stimulates erythropoiesis, which has the distinction of being the first recombinant glycoprotein produced industrially for clinical use. It is being employed on a wide scale for the treatment of anaemia in patients on haemodialysis [71]; its sales in 1991 reached $ 645 million. Another clinically important recombinant glycoprotein is the thrombolytic agent, tissue plasminogen activator (tPA), with sales of close to $ 200 million in the same year. Still, the manifold effects of carbohydrates on the stability and biological activities of glycoproteins are a source of much concern in the

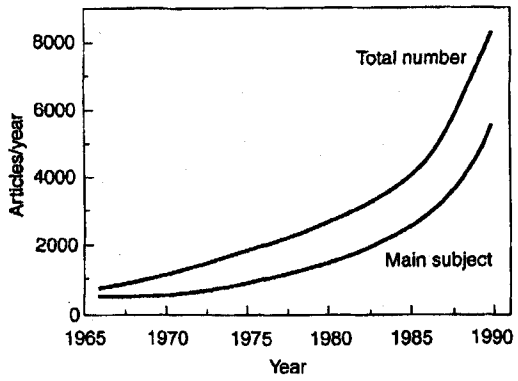



**Fig. 2. Growth of the literature on glycoproteins from 1965 to 1990, based on citations in the Medlars System.** Data obtained from Dr Elizabeth J. Van Lenten at the National Library of Medicine, National Institutes of Health, Bethesda, Maryland.

biotechnological production of pharmacologically useful glycoproteins [72–75].

As is certainly clear from the brief overview, the field of glycoproteins is expanding and growing at an extremely fast rate. This is also evidenced by the rapid growth of literature, with the number of relevant publications increasing at an almost exponential rate and currently approaching the 10000/year mark (Fig. 2). The last decade saw the birth of three specialized journals – Glycoconjugate Journal, Glycobiology and the review journal Trends in Glycoscience and Glycotechnology; the publication of the first official Nomenclature of Glycoproteins, Glycopeptides and Peptidoglycans [76], and the establishment of a computerized complex carbohydrate-structure data base (CCSD) and a corresponding management program, CarbBank [77, 78]; CarbBank now contains over 1000 structures of different N- and O-carbohydrate units found in glycoproteins (Doubert, S., personal communication). The decade also witnessed the appearance of numerous books, treatises, symposia proceedings and reviews, many of which are cited in this article. Covering the whole field in depth is an almost impossible task. We therefore chose to limit ourselves to structural and functional aspects of glycoproteins, with emphasis on the latter, and to use, as a starting point, our reviews on the subject [1, 2], as well as that by Montreuil [3], published at the beginning of the 1980s, where references to the earlier literature can be found. Also, whenever possible, we refer to review articles rather than original publications. Because of limitations of space, we shall not deal with the proteoglycans, a large group of glycoproteins with unique structural and functional properties [79–81], nor shall we discuss glycated proteins, formed under physiological conditions by nonenzymatic attachment of sugars to proteins (glycation) [82]. It should be noted, however, that much of the following discussion applies also to the sugar chains of proteoglycans, as well as to another large class of glycoconjugate, the glycolipids.

## METHODOLOGY

Structural analysis of glycoproteins is still a challenging task, not only because of the ability of sugars to form numerous isomeric compounds, even from a small number of constituents, but also because of the occurrence of glycoforms. The number of the latter can be staggering. In the protein from the scrapie prion, with a molecular mass of 27–30 kDa and two glycosylation sites, it has been calculated on the basis of the number of established oligosaccharide structures that over 400 different glycoforms can exist [83]. While isolation of individual glycoforms is not always essential for the complete structural analysis of the oligosaccharides in a given glycoprotein, it is necessary, for example, for monitoring changes in glycoform populations associated with different physiological and pathological states and for determining the structure/function relationship of individual glycoforms. The molecular masses of the individual glycoforms can be determined with great precision by electrospray mass spectrometry without fractionation when the heterogeneity is not extensive. However, in most instances only the major components are likely to be determined unambiguously [84]. In glycoproteins with a single glycan, the mass of the carbohydrate unit in each glycoform can be calculated, assuming that the exact molecular mass of the peptide moiety is known. Comparison of mass spectral data obtained before and after deglycosylation of the glycoproteins helps in determining the number of carbohydrate units and the glycosylation sites.

Separation of glycoforms is far from simple, since they often differ only in the structure of their carbohydrate units whereas differences in size or charge of the latter are infrequent. Charge differences can be artificially created, as for example through the formation of anionic borate complexes with *cis*-hydroxyl groups of sugars. Thus, the individual glycoforms of ribonuclease B have been fully resolved by capillary electrophoresis in borate buffer [85]. Affinity chromatography on lectins probably remains the most efficient method to separate glycoforms [14]. Hen ovalbumin, a glycoprotein with a single glycan but close to 20 glycoforms, has been fractionated by successive affinity chromatography on concanavalin A and wheat germ agglutinin into eight subfractions; three of these were homogenous, in that each contained only one species of carbohydrate chain [86].

Limited information on the carbohydrate structures present in a glycoprotein can be obtained after transferring it to a membrane, either directly when available as a homogenous preparation ('dot blot') or after separation by gel electrophoresis ('Western blot'); the blots are then probed with various lectins or with monoclonal antibodies prepared against specific oligosaccharide structures, both before and after treatment of the membranes with glycosidases, such as sialidase or $\beta$-galactosidase [87]. This is a fast technique, that requires only small amounts of material (from a few nanograms to about a microgram of glycoprotein), as illustrated in a recent study of glycosylation changes in IgG [88].

The complete analysis of a glycoprotein should provide information on the primary structure of each carbohydrate unit in conjunction with the nature and position of non-carbohydrate substituents, as well as on the location of the glycosylation site(s) in the polypeptide chain and the pattern of heterogeneity of each glycosylation site. Such analysis generally requires the release and subsequent recovery of the carbohydrate moieties, either in free form or as glycopeptides (Table 2) and their fractionation (Table 3) into pure, homogenous compounds [89–91]. Proteolytic degradation of glycoproteins almost always yields a mixture of glycopeptides with intact carbohydrate chains; it is the only method that, in cases of multiple glycosylation sites, allows the assignment of a given structure to a particular site. A serious drawback is that the glycopeptides are very often heterogeneous, not only in their carbohydrate moieties, but in their peptide moiety as well. Hydrazinolysis is the most common chemical

**Table 2. Release of glycans from glycoproteins.**

| Released form | Method | |
|---|---|---|
| | chemical | enzymatic |
| Oligosaccharides | hydrazinolysis[a] | treatment with endoglycosidases |
| | trifluoroacetolysis[a] | treatment with N-glycanase[a] |
| | alkaline borohydride treatment[b] | |
| Glycopeptides | | proteolysis |

[a] Applicable only for N-glycans.
[b] Used mainly for O-glycans.

**Table 3. Separation of released glycans.** Gel filtration and chromatography are performed in both conventional and HPLC modes.

| Method | Medium |
|---|---|
| Gel filtration | Sephadex<br>Sepharose<br>Bio-Gel |
| Chromatography | immobilized lectins<br>anion exchangers<br>chemically modified silica |
| Electrophoresis | paper<br>capillary zone<br>polyacrylamide gel |

procedure for the cleavage of the N-glycosidic linkage. It is, however, not absolutely specific and may cleave O-chains if they are present in the glycoprotein studied. In addition, since hydrazine attacks amide bonds indiscriminately, the oligosaccharide is de-*N*-acetylated during the procedure; this can be rectified by re-*N*-acetylation. Hydrazinolysis may also result in the removal of certain non-carbohydrate substituents. O-Glycosidic linkages to serine and threonine are routinely split off by alkaline treatment in the presence of borohydride. The reaction is, however, not quantitative: it leads to partial modification of the reducing end and to loss of some non-carbohydrate substituents.

With the discovery of endoglycosidases, the 'restriction enzymes' of the carbohydrate world, it became possible to release oligosaccharides *en bloc* from denatured, and sometimes from native, glycoproteins and from glycopeptides [92, 93]. Two classes of enzyme are available for the liberation of asparagine-linked chains, namely the endo-$\beta$-*N*-acetylglucosaminidases (e.g. endo-H and endo-F), which hydrolyze the glycosidic bond between the two *N*-acetylglucosamine residues in the *N*,*N*′-diacetylchitobiose unit next to asparagine (thus leaving a single *N*-acetylglucosamine on the protein), and the N-glycanases [for peptide-$N^4$-(*N*-acetyl-$\beta$-glucosaminyl)-asparagine amidases], which hydrolyze the GlcNAc-Asn linkage. The various endo-$\beta$-*N*-acetylglucosaminidases differ in their substrate specificities; for example, endo-H can hydrolyze most oligomannose and hybrid-type sugar chains, while endo-F cleaves also complex-type chains. The possibility of enzymatic removal of O-chains with the aid of endo-$\alpha$-*N*-acetylgalactosaminidase (O-glycanase) is limited by the very strict glycon specificity of the enzyme, which recognizes only the disaccharide Gal$\beta$3GalNAc and, as recently shown [94], the trisaccharide Fuc$\alpha$2Gal$\beta$3GalNAc. Endo-$\beta$-galactosidase, an enzyme that hydrolyses internal $\beta$-galactosyl linkages in poly(*N*-acetyllactosamine) chains, played a crucial role in studies of these glycans.

Separation of the complex mixtures of glycopeptides or oligosaccharides into homogeneous components presents a daunting task, primarily because of the very nature of such compounds. Most monosaccharide consituents of glycoproteins are closely related chemically, with similar ionic properties and frequently identical molecular masses, so that the resultant oligosaccharides are often chemically alike. Another reason is the limited amounts of material available in most cases; this can be overcome by radio-labeling the liberated oligosaccharides at their reducing termini (generally with tritiated borohydride) or by derivatization with ultraviolet-absorbing or fluorescent probes prior to fractionation, in order to facilitate monitoring [70]. Glycopeptides can be labeled at their amino groups, e.g. by radioactive acetic anhydride. Techniques available for the separation of oligosaccharides and glycopeptides are listed in Table 3.

Oligomers, consisting of 25–30 monosaccharides, can be separated by gel filtration (e.g., on Bio-Gel P-4) in both analytical and preparative modes [95]. Ion-exchange chromatography is useful as a preparative technique for separation of charged oligosaccharides that contain sialic acids, uronic acids, or charged substituents (e.g. phosphate or sulfate). In a recently developed, powerful technique, the uncharged hydroxyl groups of the sugars are transformed at high pH to negatively charged oxygen ions and the products separated by high-performance liquid chromatography on special ion-exchange columns [96]. Pulsed amperometric detection of the sugars in column effluents greatly increases the sensitivity of the procedure. Neutral oligosaccharides can be separated as borate complexes by ion-exchange chromatography or paper electrophoresis or, after tagging with a charged reagent, by gradient polyacrylamide gel electrophoresis. By choosing a highly fluorescent label (e.g. 2-aminoacridone), picomolar amounts of a saccharide can be detected [97]. An impressive increase in sensitivity has been achieved by the recent introduction of the reagent 3-(4-carboxybenzoyl)-2-quinolinecarboxaldehyde which has a detection limit in the low attomolar ($10^{-18}$) range [98]. This reagent, in combination with capillary electrophoresis, allows one to separate carbohydrate units derived from as little as subpicograms of a glycoprotein. Another possibility is to couple oligosaccharides released from glycoproteins to a suitable lipid, generating neoglycolipids, which are easily fractionated, for example by thin-layer chromatography. This allows concomitant microsequencing of the oligosaccharides, since the neoglycolipids have excellent ionisation properties in mass spectrometry [99].

Similar to its use for the fractionation of glycoproteins, affinity chromatography on immobilized lectins permits the effective separation of oligosaccharides and glycopeptides and is widely employed for this purpose [13, 14, 100]. The method is used both in its conventional mode and in high-performance liquid chromatography with increased speed and improved resolution [101]. Moreover, from 'fingerprints' obtained by lectin chromatography, structural features can be attributed to the individual oligosaccharides. Pea, lentil and *Vicia faba* lectins have long been used to separate glycopeptides with L-fucose linked $\alpha$-1,6 to the innermost *N*-acetylglucosamine from those that lack this residue. With the aid of the lectins from *Sambucus nigra* and *Maackia amurensis*, it is possible to distinguish between glycoconjugates with $\alpha$-2,3- or $\alpha$-2,6-linked sialic acid [102]. By serial affinity chro-

**Table 4. Methods used for structural elucidation of oligosaccharides.**

| Type of method | Method |
|---|---|
| Chemical | Smith degradation<br>acetolysis[a]<br>methylation analysis[b] |
| Physicochemical | mass spectrometry (fast-atom-bombardment, electron impact, etc.)<br>$^1$H-NMR and $^{13}$C-NMR |
| Enzymic | sequential degradation with exoglycosidases |

[a] Applicable for oligomannosides.
[b] Done in conjunction with GLC and often also with mass spectrometry.

matography on immobilized lectins with different specificities, a mixture of oligosaccharides (or glycopeptides) can thus be fractionated into individual constituents. This is illustrated by the separation of 16 different oligosaccharide variants of a basic di-branched structure, isolated from human IgG, into single components by sequential chromatography on columns of lectins from *Aleuria aurantia*, *Ricinus communis*, *Phaseolus vulgaris* and of concanavalin A [103]. Affinity chromatography on the latter lectin and on lectins from *P. vulgaris*, *Datura stramonium*, *R. communis* and *Ulex europaeus* of desialylated N-oligosaccharides from human von Willebrand factor, in combination with gel filtration on Bio-Gel P-4, yielded 26 single components [104].

Once a pure homogeneous oligosaccharide, or glycopeptide, has been obtained, complete structural elucidation requires, as a rule, the application of several methods in concert [105] (Table 4). This is particularly true when dealing with novel structures. Although a wide range of refined physicochemical methods are now available, simple chemical techniques, such as Smith degradation and acetolysis, are still useful. Digestion with exoglycosidases provides information on the identity, anomeric configuration and, with certain of these enzymes, also on linkage positions. For instance, $\beta$-galactosidase from *Diplococcus pneumoniae* cleaves Gal$\beta$4GlcNAc, but not Gal$\beta$3GlcNAc; $\beta$-*N*-acetylhexosaminidase from the same source recognizes only GlcNAc$\beta$2Man, out of five different *N*-acetylglucosamine linkages commonly found in glycoproteins. Another example is $\alpha$-fucosidase I from almond emulsin, which splits Fuc$\alpha$3(or 4)GlcNAc, but not Fuc$\alpha$6GlcNAc. Sialidase from Newcastle disease virus cleaves the NeuAc$\alpha$2,3Gal linkage much faster than the NeuAc$\alpha$2,6Gal linkage. A complete sequence is obtainable by stepwise treatment of an oligosaccharide with a series of glycosidases [5]. Since it is necessary to isolate the reaction products, and to characterize the released monosaccharide at each step, enzymatic degradation has to be carried out in conjunction with a chromatographic method, such as gel filtration or ion-exchange chromatography. The method has recently been adopted for fast sequencing [106].

Fast-atom-bombardment ionization [107–111] allows rapid screening of glycoproteins to ascertain the types of glycans present and the degree of heterogeneity. In conjunction with chemical methods, such as permethylation, it provides information on complete sequences of up to 15 sugar residues, the number and type of modifying groups present (acetyl, phosphate, sulfate, etc.) and the location of modified residues in the carbohydrate sequence.

The most useful technique is still $^1$H-NMR spectroscopy [112, 113] in its different forms, occasionally supplemented by $^{13}$C-NMR [114]. Although it requires larger amounts of material than certain other methods, such as fast-atom-bombardment mass spectrometry or methylation analysis, it is non-destructive and the substance analyzed can be completely recovered. The availability of extensive libraries of spectra of reference compounds permits the unequivocal establishment of already known structures and the determination of new structures. Moreover, NMR can provide information on oligosaccharide conformation and flexibility in solution, which is important in understanding the mode of carbohydrate binding to proteins.

Interesting developments are under way to provide automated machinery for glycan analysis. The arrival of commercial instruments for releasing, profiling and sequencing oligosaccharides will undoubtedly produce the same sort of revolution in glycobiology as did the automatic sequencer in protein chemistry.

## STRUCTURE

As pointed out, during the last decade there has been a large increase in the number of known sugar constituents and carbohydrate–peptide linking groups of glycoproteins, and a myriad of novel glycan structures have been characterized. Some sugars and linkages have been shown to be more widespread than previously thought.

### Monosaccharide constituents

Sulfated sugars in N-glycans proved to be much more common than heretofore suspected, especially as 4-$SO_4$-GalNAc, for example in certain pituitary hormones [115, 116] and in viruses. Galactose sulfated at 3-O was found in thyroglobulins from various sources [117, 118] and in respiratory mucins of a cystic fibrosis patient [119]; thyroglobulin also contains 6-sulfated *N*-acetylglucosamine [117, 118]. 4- or 6-Sulfated mannose occurs in ovalbumin [120] and in the slime mold (*Dictyostelium discoideum*) [121], respectively. 6-Deoxyaltrose, discovered in salmonid fish eggs, is the only 6-deoxy sugar, beside L-fucose, found in glycoproteins [122]. Furanose forms of monosaccharides are usually ignored by biochemists, even though it has been known for a time that arabinofuranose is a common constituent of O-glycans of plant glycoproteins (e.g. tomato and potato lectins) and that ribose and deoxyribose in nucleic acids are present as furanosides. More recently, galactofuranose has been shown to be present in a number of glycoproteins of bacteria [9–11] and protozoa [123, 124]. These and other examples of rare constituents are given in Table 5.

The list of sialic acids [136, 137] has continued to swell, their number now being close to 35 [138, 139]. They differ not only in the substituent on the amino group (acetyl or glycoloyl), but also in the number (up to three), position (4, 7, 8 and 9) and nature (acetyl, lactoyl and methyl) of substituents on the hydroxyl groups of neuraminic acid. A recent addition is 4,6-anhydro-*N*-acetylneuraminic, acid found in edible bird's nest substance [134]. It is generally stated that sialic acid is not found in insects; conclusive evidence has now been presented for its presence in *Drosophila melanogaster* through all stages of development [140]. *N*-Glycoloylneuraminic acid (NeuGc), once believed to be confined to pigs and horses, has now been found in most animals, except

**Table 5. Rare monosaccharide constituents of glycoproteins.**

| Monosaccharide | Source and references |
|---|---|
| 2-Acetamido-4-amino-2,4,6-trideoxyglucose | *Clostridium symbiosum* [10] |
| 6-Deoxyaltrose | salmonid fish eggs [122] |
| 3-Deoxy-D-*glycero*-*galacto*-nonulosonic (Kdn) | salmonid fish eggs [125] |
| 2,3-Diacetamido-2,3-dideoxymannuronic acid | *Bacillus stearothermopilus* [10] |
| Fuc2Me | nematode [126] |
| Galactofuranose | bacteria [11, 127]; trypanosoma [123]; yeasts [128] |
| Gal3Me | snail [129] |
| Gal4Me | nematode [126] |
| Gal6Me | alga [130] |
| Gal3$SO_3^-$ | thyroglobulin [117, 118]; mucins in cystic fibrosis [119] |
| GalNAc4$SO_3^-$ | pituitary glycohormones [115] |
| Glc3Me | *Methanothermus fervidus* [131] |
| GlcNAc3Me | *Clostridium thermocellum* [127] |
| GlcNAc6$SO_3^-$ | thyroglobulin [117, 118] |
| Gulose | alga [132] |
| Man3Me | snail [129] |
| Man4$SO_3^-$ | ovalbumin [120] |
| Man6$SO_3^-$ | ovalbumin [120]; slime mold [121] |
| Man6$PO_3^-$Me | slime mold [133] |
| ManNAc | *Clostridium symbiosum* [10] |
| 4,8-Anhydro-NeuAc | edible bird's nest [134] |
| NeuGc8Me9Ac | starfish [135] |
| NeuGc7,9$Ac_2$8Me | starfish [135] |

adult humans and birds. Using immunological methods, tiny amounts of this sialic acid have been detected in antigenic glycoproteins of some human tumors [141, 142]. The existence of these antigens has raised the possibility that the gene responsible for the synthesis of *N*-glycoloylneuraminic acid (by hydroxylation of CMP-NeuAc to CMP-NeuGc) is suppressed under normal conditions in humans, but may be induced in the course of oncogenesis. A novel type of compound, classified with the sialic acids (although not an N- or O-substituted neuraminic acid in the strict sense), is 3-deoxy-D-*glycero*-D-*galacto*-nonulosonic acid (also known as 2-keto-3-deoxy-nonulosonic acid, KDN or Kdn), in which the 5-amino group of neuraminic acid, has been replaced by a hydroxyl group. It is quite abundant in glycoproteins of salmonid fish [125] and in batracians [143].

## Carbohydrate-peptide linkages

Until recently, the only known N-linking group was *N*-acetylglucosaminyl-asparagine (GlcNAc-Asn), with the asparagine as part of the consensus sequence (or sequon) Asn-Xaa-Ser/Thr, where Xaa may be any amino acid, with the possible exception of proline [13]. It was also established that the occurrence of this sequence is not a sufficient condition for *N*-glycosylation. During recent years, several new asparagine-linked monosaccharides have been discovered, mainly in bacterial glycoproteins, namely glucose (both in $\alpha$- and $\beta$-linkage), $\beta$-*N*-acetylgalactosamine and L-rhamnose (Fig. 1). Whenever carefully studied, the asparagine in the novel linkages, too, is part of the consensus sequence mentioned above. Perhaps the only exception is nephritogenoside, in which glucose is bound in an $\alpha$-linkage to the amino-terminal sequence Asn-Pro-Leu [144].

Tyrosine has been added to the list of O-linked amino acids with the unequivocal identification of Glc$\alpha$-Tyr in glycogenin [145], the primer for glycogen synthesis [146, 147], and of Glc$\beta$-Tyr in the glycoprotein of the crystalline surface layer (termed the S layer) of *Clostridium thermohydrosulfuricum* [148]. The Gal$\alpha$-Ser linkage, previously described in plant glycoproteins (e.g. extensin and potato lectin), is also present in the glycoproteins from cellulosomes, an extracellular complex of different cellulases, produced by cellulolytic bacteria *Clostridium thermocellum* and *Bacteroides cellulosolvens* [11]. Glucose bound in a $\beta$-linkage to serine has been found in the bovine blood clotting factor IX [149]. A very unusual case is that of proteinase I from *D. discoideum*, which contains GlcNAc-1-*P* in diester linkage to serine [150].

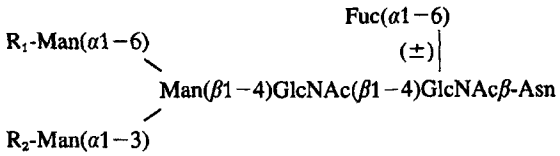
Still, the most widely occurring novel O-glycosidic linkage is probably that between *N*-acetylglucosamine and the hydroxyl group of serine/threonine [8]. Quite unusually, no other sugars are attached to the O-linked *N*-acetylglucosamine. The O-GlcNAc linking group is found predominantly in intracellular proteins exposed to the nuclear and cytoplasmic compartments. These include components of the nuclear pore, endoplasmic reticular membrane and cytoskeleton, numerous chromatin proteins, including transcription factors, as well as viral proteins.

A different kind of protein modification, involving the formation of a carbohydrate–peptide bond, is ADP-ribosylation [151]. It leads to the covalent linkage of the 3-hydroxyl of ribose with glutamic (or aspartic) acid, arginine or cysteine, and serves to attach monomeric or polymeric ADP-ribosyl residues to various proteins in different cellular compartments. Discovered about 20 years ago, ADP-ribosylation is now known to be widespread and leading to modulation of divergent processes such as DNA repair, differentiation, transmembrane signaling, carcinogenesis and bacterial nitrogen fixation.

## Oligosaccharides

The tremendous structural diversity of asparagine-linked oligosaccharides derives from variations in the number, composition and sequence of substituents attached to a pentasaccharide core, Man$\alpha$3(Man$\alpha$6)Man$\beta$4GlcNAc$\beta$4GlcNAc, common to virtually all members of this class. To the core different sugars may be attached, as well as up to five different branches (or antennae) that may differ in structure and size [3, 5]. In most plant glycoproteins, a xylose residue is attached in $\beta$-1,2 linkage to the $\beta$-linked mannose of the core, and L-fucose linked $\alpha$-1,3 (and not $\alpha$-1,6 as frequently found in animal glycoproteins) to the innermost *N*-acetylglucosamine of the core [152, 153]. The same structure is present in neural tissue of *Drosophila* and other insects [154], while the $\beta$-1,2-linked xylose (but not the $\alpha$-1,3-linked L-fucose) is found in molluscan hemocyanins [129]. A difucosylated N-glycan, with both L-fucose residues linked to the asparagine-bound *N*-acetylglucosamine, i.e. Fuc$\alpha$6(Fuc$\alpha$3)GlcNAc, was isolated from a honey bee venom glycoprotein [155]. Branches that contain sulfated sugars are generally not sialylated. An unusual structure, containing both sulfate and sialic acid on the same branch, NeuAc$\alpha$2,6Gal$\beta$4Glc($SO_4$)NAc$\beta$2 is present in thyroglobulin [118], while in recombinant tPA, expressed in mouse epithelial cells, both substituents were found attached to the same galactose residue [156]. The structure Neu5Gc8Me-$\alpha$2,5-Neu5Gc8Me, in which the glycosidic linkage is through the hydroxyl of the *N*-glycoloyl

**Table 6. Some uncommon N-glycan structures.** Structures in A are linked to the common Asn-linked pentasaccharide core.




Only branches containing unusual structure are shown. In B, modifications of the pentasaccharide core are shown.

| Structure | R | Source | Reference |
|---|---|---|---|
| A | | | |
| GalNAc(β1−4)GlcNAc(β1−2)<br>\|<br>NeuAc(α2−3/6) | $R_1$ and $R_2$ | Bowes melanoma tPA<br>lactotransferrin<br>snake venom batroxobin | [157]<br>[158]<br>[159] |
| NeuAc(α2−6)Gal(β1−3)GlcNAc(β1−4)<br>\|<br>NeuAc(α2−3) | $R_1$ and $R_2$ | fetuin | [160, 161] |
| Gal(β1−4)GlcNAc(β1−2)<br>\| \|6<br>NeuAc(α2−6) $SO_4$ | $R_2$ | thyroglobulin | [118] |
| NeuAc(α2−3)Gal(β1−4)GlcNAc(β1−2)<br>\|6<br>$SO_4$ | $R_1$ | recombinant tPA | [156] |
| Gal(β1−4)GlcNAc(β1−6/4)<br>\| \|<br>Fuc(α1−2) Fuc(α1−3) | $R_1/R_2$ | human γ-seminoprotein | [162] |
| GalNAc(β1−4)GlcNAc(β1−2) | $R_1$ and $R_2$ | *Schistosoma mansoni*<br>human kallidinogenase | [163]<br>[164] |
| GalNAc(β1−4)GlcNAc(β1−6)<br>GalNAc(β1−4)GlcNAc(β1−2) | $R_1$ | human kallidinogenase | [164] |
| Gal(β1−3)GalNAc(β1−4)GlcNAc(β1−2)<br>\|3<br>OMe | $R_1$ and $R_2$ | hemocyanin of *Lymnaea stagnalis* | [129] |
| Galf Man(α1−2)Man(α1−6)<br>Galf Man(α1−2)Man(α1−3) | $R_1$ | *Leptomonas samueli* | [123] |
| Galf Man(α1−2)Manα1−2 | $R_2$ | *L. samueli* | [123] |
| Glc(α1−3)Man(α1−2)Man(α1−6)<br>Man(α1−3) | $R_1$ | hen egg-yolk antibody | [165] |
| Glc(α1−3)Man(α1−2)Man1−2 | $R_2$ | hen egg yolk antibody | [165] |
| Glc(α1−3)Glc(α1−3)Man(α1−2)Man(α1−2) | $R_2$ | starfish egg | [166] |
| B | | | |
| Fuc(α1−6)<br>\|<br>−4GlcNAc-Asn<br>\|<br>Fuc(α1−3) | | honeybee venom phospholipase | [155] |
| Man(α1−6)<br>Man(β1−4)GlcNAc(β1−4)GlcNAc-Asn | | interleukin-6 | [167] |
| Man(α1−6)<br>Man(β1−4)GlcNAc(β1−4)GlcNAc-Asn<br>Man(α1−3)<br>\|3<br>OMe | | hemocyanin of *L. stagnalis* | [129] |

**Table 7. Some uncommon O-glycan structures.**

| Structure | Source | Reference |
|---|---|---|
| *Linked directly to serine or threonine* | | |
| Xyl(α1−3)Xyl(α1−3)Glc[a] | blood clotting factor IX | [149] |
| Gal(α1−3)GalNAc | embryonal carcinoma cells | [181] |
| NeuAc(α2−3)Gal(β1−3)GalNAc<br>\|<br>NeuAc(α2−8)NeuAc(α2−6) | glycophorin | [182] |
| 4-OMe-Gal(β1−3)GalNAc<br>\|<br>2-OMe-Fuc(α1−2) | nematode | [126] |
| $SO_4$-6Gal(β1−3)GalNAc<br>\|<br>NeuAc(α2−6) | mucin of patient with cystic fibrosis | [183] |
| NeuAc(α2−6)Gal(β1−3)GalNAc<br>\|<br>$SO_4$-6GlcNAc(β1−6) | mucin of patient with cystic fibrosis | [183] |
| *Peripheral oligosaccharides* | | |
| Gal(β1−4)GlcNAc(β1−3)Gal(β1−4)-R[b]<br>\| \| \|<br>Fuc(α1−2) \| Fuc(α1−2)<br>Fuc(α1−3) | mucin of patient with bronchiectasis | [184] |

[a] Found only as serine-linked.
[b] R = GlcNAc(β1−6)[GlcNAc(β1−3)]GalNAc-Ser/Thr or GlcNAc(β1−6)[Fuc(α1−2)Gal(β1−3)]GalNAc-Ser/Thr.

moiety, has been isolated from the starfish *Asterias rubens* [135]; its origin, whether glycoprotein or glycolipid, is, however, not known. These and other uncommon structures are listed in Table 6.

The branches of many N-oligosaccharides in animal cells [168], and as shown recently also in a glycoprotein from the protozoan *Trypanosoma brucei* [169], contain poly-(*N*-acetyllactosamine), a polymer of β-1,3-linked *N*-acetyllactosamine, Galβ4GlcNAcβ, with a degree of polymerization as high as 50. Due to the branch specificity of β-1,3-*N*-acetylglucosaminyltransferase, the enzyme that initiates the formation of poly-(*N*-acetyllactosamine) chains and elongates them [44], such chains are predominant on the Manα6Manβ branch of the core pentasaccharide. The poly-(*N*-acetyllactosamine) chains, with their repeating disaccharide and frequent substitution of their galactose residues by GlcNAcβ1,6 branches, can form a tremendous number of different structures. Among others, they serve as backbone structures for the ABO, I/i and Lewis blood group antigenic determinants on human erythrocyte membrane glycoproteins (and glycolipids) [170].

Another type of polymer, confined largely to glycoproteins of neural origin and of fish eggs, are the polysialic acids [171, 172]. They may be N-linked, via the pentasaccharide core, or O-linked (see below). In neural glycoproteins, where polysialic acid is part of a class of adhesion molecules known as neural cell-adhesion molecules (N-CAMs) [173], it is made up exclusively of α-2,8-linked *N*-acetylneuraminic acid, with a degree of polymerization as high as 55 [174]. In contrast, polysialic acids from fish egg glycoproteins exhibit an extraordinary variation in chemical structure, due to the presence of both *N*-acetyl- and *N*-glycoloylneuraminic acid, of different linkages and various substitutions on their hydroxyls. Polymers of α-2,8, α-2,9 and α-2,8/α-2,9-linked *N*-acetylneuraminic acid occur in the capsule of *Escherichia coli* and certain other Gram-negative bacteria [172].

Besides those mentioned above, the only glycoprotein reported to contain polysialic acid is the voltage-sensitive sodium channel from *Electrophorus electricus* electroplax [175] and from adult rat brain [176]. No information is available regarding the length of the sialic acid chains or their function.

Saccharides, bound to the proteins via the GalNAc-Ser/Thr linkage, represent a large and ubiquitous family of O-glycans [177]. They vary in size from a monomer to oligomers of up to 20 constituents. Monomeric and dimeric units are quite rare, one example being *N*-acetylgalactosamine and NeuAcα2,6GalNAc in glycophorin of individuals of the Tn phenotype [178] and on a variety of human cancer cells. Larger O-glycans are common in the mucins, a group of glycoproteins found in epithelial secretions and as membrane constituents in epithelial and certain tumor cells [179, 180]. Mucins are giant molecules, of 4−6 MDa, each of which may contain up to several hundred oligosaccharide chains. A typical constituent of the O-glycans is *N*-acetylgalactosamine, accompanied by galactose, *N*-acetylglucosamine, L-fucose, various sialic acids and sulfate. The structure $SO_4$-3Galβ4GlcNAcβ6GalNAc-Ser/Thr was found in mucins isolated from sputum of patients with cystic fibrosis [119] (Table 7). Tracheobronchial mucus glycoproteins from such patients also contain oligosaccharides with both a sulfate ester and a sialic acid residue, sometimes on the same branch [183], e.g. $SO_4$-Galβ3(NeuAcα2,6)GalNAc. Poly-(*N*-acetyllactosamine) chains carrying several fucose residues, e.g. Fucα2Galβ4(Fucα3)GlcNAcβ3(Fucα2)Galβ4-R, were isolated from respiratory mucins of a patient suffering from bronchiectasis [184]. Some mucins bear different blood type determinants and exhibit a striking structural heterogeneity.

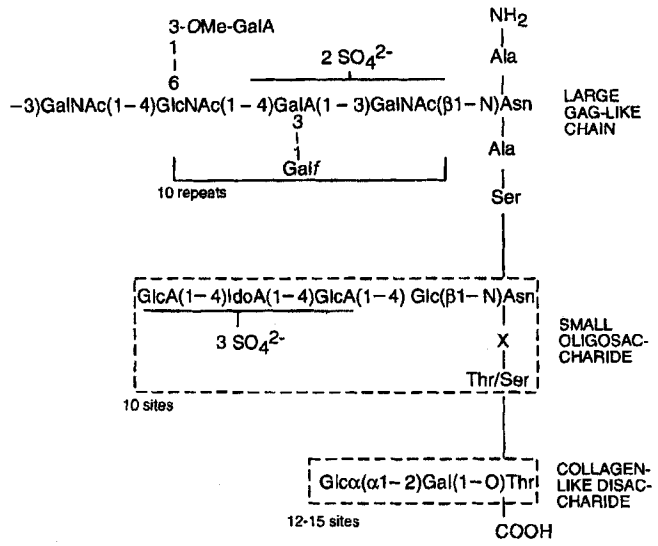



**Fig. 3. Carbohydrate chains of cell surface glycoprotein of halobacteria.** (From [185].) GAG = glycosaminoglycan.

O-linked polysialic acid in glycoproteins from the eggs of rainbow trout, in which *N*-glycoloylneuraminic acid accounts for about 60% of the mass, have the structure Gal$\beta$3[(NeuGc$\alpha$2,8)$_n$NeuGc$\alpha$2,6]GalNAc.

As mentioned, the carbohydrates of bacterial glycoproteins are unusual, not only with respect to the carbohydrate–peptide linking groups, but also to the overall structure, which is quite diverse (Fig. 3).

Several rare structures were identified in free oligosaccharides. Thus, NeuAc$\alpha$2,6Man$\beta$4GlcNAc has been isolated from the urine of a patient with $\beta$-mannosidosis, a genetic disease of glycoprotein catabolism that has been described in the 1980s [186]. This trisaccharide is probably the product of the enzymatic transfer of *N*-acetylneuraminic acid to Man$\beta$4GlcNAc that accumulates in such patients. In bovine colostrum, a trisaccharide with the structure NeuAc$\alpha$2,6Gal$\beta$4GlcNAc-6-*P* is present [187]. The biosynthetic origin of this compound is not known. It is possible that structures of the types described occur in glycoproteins, but no evidence for this is available.

### The glycosyl-phosphatidylinositol anchor

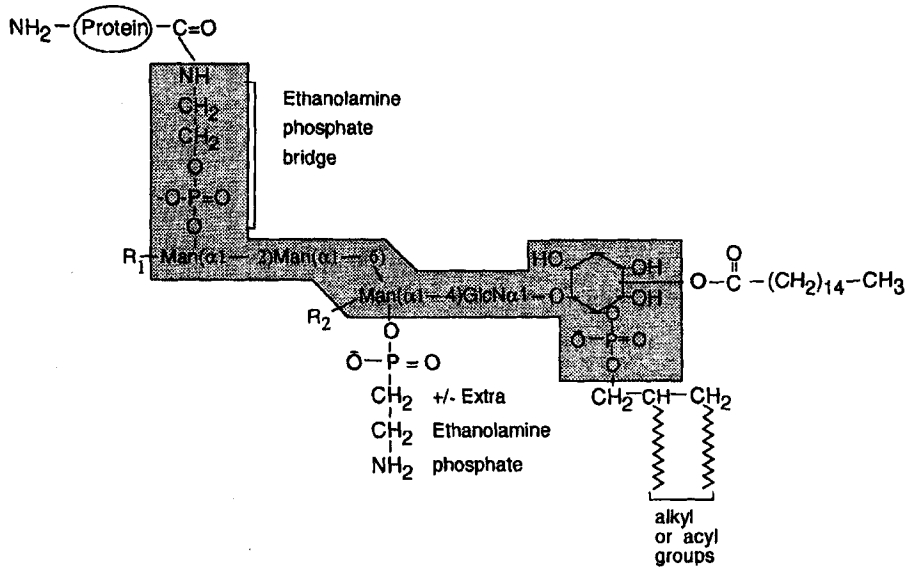
Although direct documentation for the occurrence of GlPtdIns anchors was obtained only in 1985 [15], over 100 proteins with such anchors have now been identified. They were found in organisms representing most stages of eukaryotic evolution, from protozoa, yeast and slime mold to *Drosophila* and man, but none has been reported in prokaryotes and only very recently has evidence been obtained for their presence in plants (Gibbs, B. J. and Thomas-Oates, J., personal communication). The parasitic protozoa also express large amounts of structurally related glycolipids which are not attached to proteins, such as lipophosphoglycans.

Identification of GlPtdIns anchors is often based on their susceptibility to purified bacterial phosphatidylinositol-specific phospholipase C, which leads to the release of the anchored protein from the cell surface. In a few instances, the complete structure of the GlPtdIns has been established [16–18]. These include the GlPtdIns from the rat brain glycoprotein Thy-1, human erythrocyte acetylcholinesterase, the variant surface glycoprotein (VSG) and the 1G7 antigen of the parasitic protozoan *T. brucei* [188]. All these structures share a common tetrasaccharide core, Man$\alpha$2Man$\alpha$6Man$\alpha$4GlcN$\alpha$, the most unusual feature of which is that the amine of glucosamine is free, and not substituted, e.g. by an acetyl group as in most glycoproteins. In fact, the non-acetylated glucosamine is a diagnostic feature of GlPtdIns. The tetrasaccharide is bound via the 6-hydroxyl group of mannose at its nonreducing end to ethanolamine phosphate, which in turn is attached by an amide linkage to the $\alpha$-carboxyl of the terminal amino acid of the protein. The reducing end of the tetrasaccharide is glycosidically linked to an inositol phospholipid which is embedded in the lipid layer of the cell surface membrane, thus mediating the binding of the protein to the membrane. The anchors differ in the nature and number of additional carbohydrate and ethanolamine phosphate moieties linked to the tetrasaccharide core. For example, the *T. brucei* anchor has a side chain, composed of $\alpha$-linked galactose residues, attached to C3 of the glucosamine-bound mannose, whereas the rat brain Thy-1 has a $\beta$-*N*-acetylgalactosamine residue and an ethanolamine phosphate moiety linked to C4 and C2, respectively, of the same mannose, and an additional mannose linked to the terminal, nonreducing mannose of the conserved core (Fig. 4). Furthermore, there is structural heterogeneity within the GlPtdIns anchors isolated from the same tissue or cell type, as well as tissue-specific differences. Thus, about 30% of the rat brain Thy-1 lack the extra mannose, which is completely absent in rat thymocyte Thy-1; most of the latter is also devoid of the *N*-acetylgalactosamine. In the *T. brucei* surface glycoprotein, the number of galactose residues in the side chain varies between 0 and 8. There is also considerable variation in the fatty acyl or fatty alkyl groups of the phospholipid; the variant surface glycoprotein of *T. brucei* is unusual in containing exclusively myristate.

## CONFORMATION OF OLIGOSACCHARIDES

Oligosaccharides are flexible molecules that may adopt different shapes, because there is considerable freedom of rotation around the glycosidic bond, in particular to C6 of hexopyranoses [24–30]. Molecular modelling and high-resolution NMR spectrometry, often complemented by energy calculations, have shown that, in practice, rotation around some of the linkages is restricted. As a result, oligosaccharides exist in solution with regions of well-defined three-dimensional structures and relatively few degrees of freedom. This has been most clearly demonstrated for N-glycans. In the common pentasaccharide core, $Man_3(GlcNAc)_2$, rigidity is associated with the Man$\alpha$3Man$\beta$4GlcNAc$\beta$4GlcNAc fragment, while the $\alpha$-1,6-linked branch can form two rotational isomers relative to the C5-C6 bond of the $\beta$-1,4-linked mannose. The prevalence of either of the two forms depends on the type of substitution on the latter residue. In particular, attachment of a $\beta$-1,4→linked (bisecting) *N*-acetylglucosamine fixes the orientation of the $\alpha$-1,6 arm into one of the two possible conformations, folded back towards the core. In tri- or tetra-antennary oligosaccharide chains, with GlcNAc$\beta$-1,6 and GlcNAc$\beta$-1,2 branches linked to the same mannose residue, the $\beta$-1,6 linkage is restrained by steric interactions between the *N*-acetylglucosamine residues of the two branches.

Studies of the conformation of O-glycopeptides have focused on the molecular arrangement in the carbohydrate–




| Protein | $R_1$ | $R_2$ | Extra ethanolamine phosphate | Palmitoylation |
|---|---|---|---|---|
| Human erythrocyte AChE | – | – | + | + |
| *Leishmania major* PSP | – | – | – | – |
| Rat brain Thy-1 | Man(α1–2) | GalNAc(β1–4) | + | – |
| *Trypanosoma brucei* VSG | – | ± Gal(α1–2)Gal(α1–6)<br>Gal(α1–3)<br>± Gal(α1–2) | – | – |
| *Trypanosoma cruzi* 1G7 | Man(α1–2) | – | – | ? |

**Fig. 4. Structure of the glycosyl-phosphatidylinositol (G1Ptdlns) anchor.** (Modified from [18].)

peptide linking region. The protein backbone of mucins, both soluble and membrane-associated, consists of repetitive sequences. These repeats are rich in serine and threonine residues, the potential O-glycosylation sites. They usually also contain many proline residues and other helix-breaking amino acids, resulting in molecules with an extended structure and many β-turns. This structure becomes very rigid by the addition of numerous O-glycans, resulting in mucin domains that may reach a length of several hundred nanometers.

In glycoproteins, the carbohydrates are usually on the surface of the molecule. Attempts to determine by crystallographic methods the three-dimensional structure of these units have been hampered by their inherent mobility and chemical heterogeneity. From the evidence available it is clear, however, that out of the vast ensemble of conformations accessible to exposed oligosaccharides, a particular conformation may be stabilized by interactions with the protein. IgG contains two biantennary complex oligosaccharides, each linked to Asn247 of the heavy chain; there are at least 30 structural variants of these units. In the Fc fragments of human and rabbit IgG, X-ray crystallography revealed that the α-1,6 branches of each oligosaccharide make several contacts, mainly hydrophobic, with amino acid residues on the surface of the protein [189]. In contrast, the α-1,3 branches are oriented toward each other and the *N*-acetylglucosamine residue in the outer arm of one glycan interacts with *N*-acetylglucosamine and mannose in the core of the other. Most residues of the Fc glycans are fixed and relatively immobile also in solution, as shown by experiments in which a spin label was introduced into the carbohydrate [190]. It was found that the sugar units rotate in solution together with the protein, and have no freedom of rotation of their own.

In the crystal of the glycoprotein lectin from *Erythrina corallodendron*, which consists of two identical subunits, each containing the N-linked heptasaccharide Manα6-(Manα3)(Xylβ2)Manβ4GlcNAc(Fucα3)β4GlcNAc the carbohydrate is tethered to the protein by a network of intra- and intermolecular hydrogen bonds and as a result all seven monosaccharides are seen with exceptional clarity in the electron density map [191]. The heptasaccharide projects out from the protein surface and adopts an extended conformation, with the β-1,4-linked *N*-acetylglucosamine and the terminal trimannoside unit roughly coplanar and nearly perpendicular to the mean plane of the first, N-linked *N*-acetylglucosamine (Fig. 5). The latter monosaccharide and its α-1,3-linked L-fucose are held tightly by hydrogen bonds to side-chain atoms of amino acids of the parent subunit, while the other saccharides form hydrogen bonds with main-chain and side-chain atoms of amino acids of another, symmetry related, subunit. The 'frozen' conformation of the heptasaccharide when bound to the protein is well in the range of conformations found for similar or identical oligosaccharides in solution, showing that it is not forced by the protein into an unusual conformation, Despite their relative proximity, there are no direct interactions between the carbohydrate

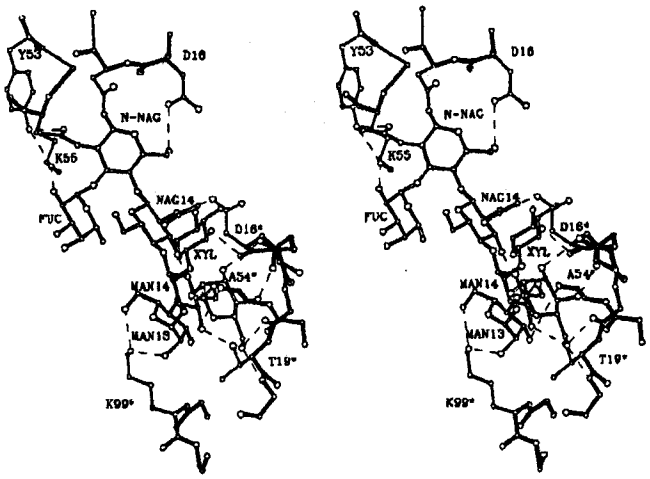



**Fig. 5. The network of hydrogen bonds, stabilizing the N-linked carbohydrate in *Erythrina corallodendron* lectin (stereo).** One-letter code and sequence number for shown protein residues; asterisks indicate symmetry-related residues. Dashed lines = hydrogen bonds; empty bonds = carbohydrate; full bonds = amino acids. Water molecules omitted for clarity (from [191]).

chains of the two subunits, in contrast to what has been observed in IgG.

High-resolution X-ray studies of the trisaccharide Man$\alpha$-3Man$\beta$4GlcNAc in complex with the *Lathyrus ochrus* lectin revealed that it adopts an extended conformation close to the most stable one found in solution [192]. Only the $\alpha$-1,3-linked mannose establishes direct hydrogen bonds with the carbohydrate binding site of the lectin, while the other contacts are mostly mediated by sugar-water-lectin linkages. In a similar study of the complex of the same lectin with a biantennary *N*-acetyllactosamine-type octasaccharide from human lactotransferrin, GlcNAc$\beta$4Gal$\beta$2Man$\alpha$6(GlcNAc$\beta$4-Gal$\beta$2Man$\alpha$3)Man$\beta$4GlcNAc, it was again found that the conformation of the sugar is only slightly modified upon binding [193]. Besides the contacts between the lectin and the $\alpha$-1,3-linked mannose in the binding site, the complex is stabilized by numerous hydrogen bonds, many of them involving water, and by van der Waals interactions with most of the other monosaccharide residues of the octasaccharide. Comparison of the Man$\alpha$3Man$\beta$4GlcNAc trisaccharide complexed with the lectin with the corresponding part of the octasaccharide showed that they display very different conformations, except for the mannose in the combining site. Thus, the conformation of the same trisaccharide may change when it is part of a larger structure.

## FUNCTIONS

A variety of approaches are being employed in the quest to unravel the role(s) of the carbohydrate units of glycoproteins. Some of these have been mentioned earlier. They include modification of glycans by purified glycosidases and transferases, use of inhibitors of glycosylation or glycoprotein processing and of cell mutants with known defects in glycosylation and, more recently, techniques of molecular genetics. Recombinant glycoproteins can be expressed in different cells or organisms, resulting in different patterns of glycosylation. An extreme example is that of bacteria, (e.g. *E. coli*), which produce completely non-glycosylated proteins. Most insects lack the ability to sialylate glycoproteins,

**Table 8. Functions of glycoprotein glycans.**

| Type | Function |
|---|---|
| Physico-chemical | Modify solubility, electrical charge, mass, size and viscosity in solution<br>Control protein folding<br>Stabilize protein conformation<br>Confer thermal stability and protection against proteolysis |
| Biological | Regulate intracellular traffic and localization of glycoproteins<br>Determine lifetime of glycoproteins in circulation<br>Modify immunological properties<br>Modulate activity of enzymes and hormones<br>Act as cell surface receptors for lectins, antibodies, toxins, etc.<br>Participate in cell–cell interactions |

while Chinese hamster ovary (CHO) cells do not transfer sialic acid in $\alpha$-2,6 linkage, nor do they make the Gal$\alpha$3 linkage. Nucleotide-directed mutagenesis can be applied to modify glycosylation sites so that they will no longer serve as acceptors in the oligosaccharide transfer reaction. With N-glycoproteins, where the carbohydrate is attached to the Asn-Xaa-Ser/Thr sequon, modification of either the first or third amino acid will abolish glycosylation at this site. In glycoproteins with more than one carbohydrate unit, whether N-, O- or both, the sites can be systematically eliminated in various combinations, to form a panel of mutants in which the roles of each carbohydrate chain can be assessed. Finally, the use of transgenic animals, still in its infancy, may allow the study of carbohydrate function in the intact organism [194]. It should be kept in mind, however, that each of the above approaches suffers from certain limitations. Thus, inhibitors of glycosylation act indiscriminately and lack of glycosylation of other glycoproteins may indirectly affect expression or function of the glycoprotein under investigation. Similarly, in transgenic animals, deletion or insertion of a glycosyltransferase gene, for example, may affect many glycoproteins, since glycans are secondary gene products. Effects observed after elimination of a glycosylation site by mutagenesis may be due to differences in protein folding, resulting from the change in amino acid sequence and not from the absence of a glycan in the particular site.

Much of the work to be discussed deals with the roles of N-glycans. The importance of O-sugar chains in glycoproteins has been less investigated, not the least because of the scarcity of specific inhibitors of O-glycosylation. O-glycans are also less amenable to genetic manipulation since, unlike N-deglycosylation, the requirements for O-glycosylation are not obvious at the sequence level. The isolation of a mutant line of CHO cells (ldlD cells) deficient in the enzyme UDP-Gal/UDP-GalNAc 4-epimerase, which allows control over O-glycosylation by omitting or including *N*-acetylgalactosamine in the medium, has provided, for the first time, an efficient tool for the study of the function of O-oligosaccharides [195].

The overall conclusion from the various studies has been that glycans, whether N- or O-linked, perform diverse roles (Table 8) and that different glycoproteins have different requirements for carbohydrates, while in many cases, no marked effect of the carbohydrate on the properties of the