Clinical Chemistry 52:4
574–600 (2006)

Reviews

Mechanisms in Protein O-Glycan Biosynthesis and Clinical and Molecular Aspects of Protein O-Glycan Biosynthesis Defects: A Review

Suzan Wopereis,[1] Dirk J. Lefeber,[1] Éva Morava,[2] and Ron A. Wevers[1*]

Background: **Genetic diseases that affect the biosynthesis of protein O-glycans are a rapidly growing group of disorders. Because this group of disorders does not have a collective name, it is difficult to get an overview of O-glycosylation in relation to human health and disease. Many patients with an unsolved defect in N-glycosylation are found to have an abnormal O-glycosylation as well. It is becoming increasingly evident that the primary defect of these disorders is not necessarily localized in one of the glycan-specific transferases, but can likewise be found in the biosynthesis of nucleotide sugars, their transport to the endoplasmic reticulum (ER)/Golgi, and in Golgi trafficking. Already, disorders in O-glycan biosynthesis form a substantial group of genetic diseases. In view of the number of genes involved in O-glycosylation processes and the increasing scientific interest in congenital disorders of glycosylation, it is expected that the number of identified diseases in this group will grow rapidly over the coming years.**

Content: **We first discuss the biosynthesis of protein O-glycans from their building blocks to their secretion from the Golgi. Subsequently, we review 24 different genetic disorders in O-glycosylation and 10 different genetic disorders that affect both N- and O-glycosylation. The key clinical, metabolic, chemical, diagnostic, and genetic features are described. Additionally, we describe methods that can be used in clinical laboratory screening for protein O-glycosylation biosynthesis defects and their pitfalls. Finally, we introduce existing methods that might be useful for unraveling O-glycosylation defects in the future.**

The human proteome, originating from expression of the protein-coding genes of the genome, comprises ~30 000 proteins (*1*), a surprisingly low number considering that the genome of the nematode *Caenorhabditis elegans* comprises 20 000 genes (*2*). However, a higher order of complexity of protein products in humans arises from pretranslational events, such as alternative splicing, and posttranslational modifications, such as phosphorylation and glycosylation. Glycosylation, the enzymatic addition of carbohydrates to proteins or lipids, is the most common and most complex form of posttranslational modification. This is illustrated by the estimation that 1% of human

[1] Laboratory of Pediatrics and Neurology and [2] Department of Pediatrics, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands.

[*] Address correspondence to this author at: Laboratory of Pediatrics and Neurology (830), Institute of Neurology, Radboud University Nijmegen Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, The Netherlands. Fax 31-24-3540297; e-mail r.wevers@cukz.umcn.nl.

Received November 2, 2005; accepted January 24, 2006.

Previously published online at DOI: 10.1373/clinchem.2005.063040

[3] Nonstandard abbreviations: hLys, hydroxylysine; CDG, congenital disorders of glycosylation; GalNAc, *N*-acetylgalactosamine; NeuAc, *N*-acetylneuraminic acid (sialic acid); GlcNAc, *N*-acetylglucosamine; sLe^x, sialyl Lewis^x antigen; GAG, glycosaminoglycan; GlcA, glucuronic acid (or glucuronate); EGF, epidermal growth factor; TSR, thrombospondin type-1 repeat; ER, endoplasmic reticulum; GNE/MNK, UDP-GlcNAc 2 epimerase/*N*-acetylmannosamine kinase; Dol-P, dolichol phosphate; NST, nucleotide sugar transporter; CHO, Chinese hamster ovary; FUCT, GDP-Fuc transporter; β3-Gal-T, β3-galactosyltransferase; Cosmc, core 1 β3-Gal-T-specific molecular chaperone; pp-GalNAc-T, polypeptide *N*-acetylgalactosaminyltransferase; EXTL exostoses-like; COP, coatomer protein; ERGIC, endoplasmic reticulum-Golgi intermediate compartment; SNARE, soluble *N*-ethylmaleimide-sensitive fusion attachment protein receptor; COG, conserved oligomeric Golgi complex; GalNT, *N*-acetylgalactosyltransferase; FTC, familial tumoral calcinosis; B4GalT, β-1,4-galactosyltransferase; HME, hereditary multiple exostoses; MCD, macular corneal dystrophy; SED, spondyloepiphyseal dysplasia; DTDST, diastrophic dysplasia sulfate transporter; DTD, diastrophic dysplasia; ACGB1, achondrogenesis type 1B; AO-II, atelosteogenesis type II; EDM4, multiple epiphyseal dysplasia 4; PAPSS2, 3′-phosphoadenosine 5′-phosphosulfate synthase 2; APS, adenosine 5′-phosphosulfate; PAPS, 3′-phosphoadenosine 5′-phosphosulfate; WWS, Walker–Warburg syndrome; LGMD2, limb-girdle muscular dystrophy type 2; MEB, muscle–eye–brain disease; FCMD, Fukuyama-type congenital muscular dystrophy; FKRP, fukutin-related protein; MDC, congenital muscular dystrophy; LARGE, *N*-acetylglucosaminyl-like protein; hIBM, hereditary inclusion body myopathy; DMRV, distal myopathy with rimmed vacuoles; FUT, fucosyltransferase; IEF, isoelectric focusing; apoC-III, apolipoprotein C-III; and CMRD, chylomicron retention disease.

genes are required for this specific process (*3*). Furthermore, more than one half of all proteins are glycosylated, according to estimates based on the SwissProt database (*4*). In humans, protein-linked glycans can be divided into 3 categories: N-linked (linkage to the amide group of Asn), O-linked [linkage to the hydroxyl group of Ser, Thr, or hydroxylysine (hLys)[3]], and C-linked (linkage to a carboxyl group of Trp) (*5*).

Initially, the study of glycoproteins and their role in human congenital diseases focused on N-linked glycans. The diseases in this pathway have collectively been referred to as congenital disorders of glycosylation (CDG). N-Glycans share a common protein–glycan linkage and have a common biosynthetic pathway that diverges only in the late Golgi stage. Endoglycosidases are available that can cleave intact N-glycans from the protein backbone, making it relatively easy to study alterations of N-glycosylation in health and disease. In contrast, O-glycans are built on different protein glycan linkages and have extremely diverse structures; in addition, there is no endoglycosidase available for the release of intact O-glycans. However, methods for the chemical release of O-glycans have been developed and have enabled the generation of structural information for O-glycans, making it more feasible to study alterations in O-glycosylation in relation to health and disease. This review focuses on the biosynthesis of O-glycans and the human congenital disorders of O-glycosylation and their screening.

Structures of O-Linked Glycans

The O-glycosylation process produces an immense multiplicity of chemical structures. Each monosaccharide has 3 or 4 attachment sites for linkage of other sugar residues and can form a glycosidic linkage in an α or β configuration, allowing glycan structures to form branches. Glycans therefore have a larger structural diversity in contrast to other cellular macromolecules such as proteins, DNA, and RNA, which form only linear chains. Theoretically, the 9 common monosaccharides found in humans could be assembled into more than 15 million possible tetrasaccharides, all of which would be considered relatively simple glycans (*6*).

The 7 different types of O-linked glycans found in humans are summarized in Table 1. O-Linked glycans are classified on the basis of the first sugar attached to a Ser, Thr, or hLys residue of a protein. The mucin-type O-glycan, with *N*-acetylgalactosamine (GalNAc) at the reducing end, is the most common form in humans. In total, 8 mucin-type core structures can be distinguished, depending on the second sugar and its sugar linkage, of which cores 1–6 and core 8 have been described in humans (summarized in Table 2) (*7*). In addition to the 7 core structures, the Tn (GalNAcα1-Ser/Thr) and sialyl Tn [NeuAcα2–6GalNAcα1-Ser/Thr; where NeuAc is *N*-acetylneuraminic acid (sialic acid)] epitopes can be distinguished. The core structures can be further modified; for example, by the addition of an *N*-acetyllactosamine unit (Galβ1–4GlcNAc; where GlcNAc is *N*-acetylglucosamine), also seen on N-glycans. The *N*-acetyllactosamine unit may be branched by a GlcNAcβ1–6 residue or form repeating *N*-acetyllactosamine units, called poly *N*-acetyllactosamine extensions. It can also attach to the blood group determinants (A, B, and H) and the type 2 Lewis determinants [Le^x, sialyl $Lewis^x$ (sLe^x), and Le^y]. *N*-Acetyllactosamine elongations are seen mainly on core 2 O-glycans. Sugars occurring at the nonreducing termini include NeuAc, Fuc, GlcNAc, and GalNAc. GlcNAc and Gal residues can be modified at position 6 or at positions 3 and/or 6, respectively, by sulfation (*8*), and NeuAc residues can be further modified at positions 4, 7, 8, and 9 with O-acetyl ester groups (*9*). This gives rise to several hundreds of different mucin-type O-glycan structures, of which core 1 and 2 are most abundant (*7*).

Another common type of O-glycosylation with large structural diversity involves the glycosaminoglycans (GAGs). Proteoglycans are proteins containing GAG chains. GAGs are attached to a Ser residue of a protein via the linker tetrasaccharide GlcAβ1–3Galβ1–3Galβ1–4Xylβ1, except for keratan sulfate, which is linked to proteins either through N- or core 1 O-glycans. GAGs are long, unbranched polysaccharides containing a disaccharide repeat that consists of either a GalNAc or GlcNAc residue combined with a glucuronic acid (GlcA) or a Gal residue. Three different types of GAGs can be distinguished on the basis of the composition of the disaccharide repeat: (*a*) dermatan sulfate and chondroitin sulfate (GlcA + GalNAc); (*b*) heparin/heparan sulfate (GlcA + GlcNAc); and (*c*) keratan sulfate (Gal + GlcNAc). GlcA in dermatan sulfate and heparin/heparan sulfate can be epimerized to iduronate. The heterogeneity of GAGs results from variable O-

Table 1. Different types of O-linked glycans in humans.

Type of O-linked glycan	Structure and peptide linkage	Glycoprotein	Reference(s)
Mucin-type	(R)-GalNAcα1-Ser/Thr	Secreted + plasma membrane	*(8)*
GAG	(R)-GlcAβ1–3Galβ1–3Galβ1–4Xylβ1-Ser	Proteoglycans	*(216, 217)*
O-linked GlcNAc	GlcNAcβ1-Ser/Thr	Nuclear and cytoplasmic	*(218)*
O-linked Gal	Glcα1–2±Galβ1-*O*-Lys	Collagens	*(219)*
O-linked Man	NeuAcα2–3Galβ1–4GlcNAcβ1–2Manα1-Ser/Thr	α-Dystroglycan	*(16)*
O-linked Glc	Xylα1–3Xylα1–3±Glcβ1-Ser	EGF protein domains	*(220)*
O-linked Fuc	NeuAcα2–6Galβ1–4GlcNAcβ1–3±Fucα1-Ser/Thr	EGF protein domains	*(220)*
	Glcβ1–3Fucα1-Ser/Thr	TSR repeats	*(24)*

Table 2. Diversity of mucin-type O-linked glycans.

Core	Structure	Human tissue	Reference(s)
1	Galβ1–3GalNAc	Most cells and secreted proteins	*(7)*
2	Galβ1–3 (GlcNAcβ1–6)GalNAc	All blood cells	*(221)*
3	GlcNAcβ1–3GalNAc	Colon and saliva	*(222, 223)*
4	GlcNAcβ1–3 (GlcNAcβ1–6)GalNAc	Mucin-secreting cell types	*(221)*
5	GalNAcα1–3GalNAc	Meconium	*(224)*
6	GlcNAcβ1–6GalNAc	Ovarian tissue	*(225)*
7	GlcNAcα1–6GalNAc		
8	Galα1–3GalNAc	Bronchia	*(226)*

sulfation at defined locations *(10)*. An extra modification step occurs in heparin and heparan sulfate by the deacetylation and *N*-sulfation of GlcNAc residues. Regions in which the hexosamine units are acetylated remain (almost) unmodified and consist of disaccharide repeats with GlcA, whereas regions with deacetylated hexosamine units become highly sulfated and exist as disaccharide repeats with iduronate. Heparin is a highly and uniformly sulfated GAG, whereas heparan sulfate is highly sulfated only in defined blocks *(11)*.

The structures of the other 5 O-glycan types seem to show less variability, and they occur mostly in one conformation. A frequently occurring O-linked glycan is the single GlcNAc linked to nuclear and cytosolic proteins. This posttranslational modification is more analogous to phosphorylation than to classical complex O-glycosylation because it is a reversible process catalyzed by the enzymes O-GlcNAc transferase and O-GlcNAcase, respectively *(12)*, and the "normal glycosylation machinery" is not implicated *(12, 13)*.

O-Galactosyl glycans have been found only on collagen domains. Gal or Glcα1–2Gal residues are covalently linked to hLys residues found in collagens, but not all hLys residues become glycosylated. The collagen 3-dimensional structure depends on the extent of this posttranslational modification. The quantities and types of O-galactosyl glycans vary considerably not only among the different types of collagen, but also among the same collagen type from different tissues and even the same collagen type from different areas of the same type of tissue *(14, 15)*.

O-Mannosyl glycans are a less common type of protein modification, present on a limited number of glycoproteins in the brain, nerves, and skeletal muscle. The best known O-mannosylglycosylated protein is α-dystroglycan, which is a skeletal muscle extracellular matrix protein *(16)*. To date, only the NeuAcα2–3Galβ1–4GlcNAcβ1–2Man structure has been found in humans. α-Dystroglycan containing Galβ1–4(Fucα1–3)GlcNAcβ1–2Man has been found in sheep brain *(17, 18)*, and the O-mannosyl glycan HSO_3-3GlcAβ1–3Galβ1–4GlcNAcβ1–2Man has been detected in rat brain *(18, 19)*. Studies have also shown that mammalian *N*-acetylglucosaminyltransferase IX acts on the GlcNAcβ1,2-Manα1-Ser/Thr moiety, suggesting that 2,6-branched O-mannosyl glycan structures are formed in the brain *(20)*. It is therefore likely that structural diversity of O-mannosyl glycans will also be present in humans.

O-Glucosyl and O-fucosyl glycans are also rare types of protein glycosylations that have been found in the epidermal growth factor homology regions (EGF modules) of some human proteins. An EGF module is a common structural motif found in several secreted and cell-surface proteins that is often involved in mediating protein–protein interactions. The EGF repeat is typically 30–40 amino acids long and is characterized by 6 conserved Cys residues participating in 3 disulfide bridges. Glc is linked to the Ser residue in proteins in the putative consensus sequence C^1XSXPC^2 (where C^1 and C^2 are the first and second conserved cysteines of the EGF module, S is the modified Ser residue, and X can be any amino acid) *(21)*. O-Linked Glc can be further elongated with 1 or 2 α1–3 linked xyloses and is found on proteins such as human factor VII, factor IX, and protein Z *(22, 23)*. All O-fucosylated glycoproteins are modified with a single O-linked Fuc residue (e.g., urinary-type plasminogen activator, tissue-type plasminogen activator, and coagulation factors VII and XII) except for coagulation factor IX, which contains O-linked Fuc that is elongated to the tetrasaccharide NeuAcα2–6Galβ1–4GlcNAcβ1–3Fucα1-Ser/Thr. Most O-Fuc modifications on EGF repeats are found on the consensus site $C^2X_{3-5}S/TC^3$ (where C^2 and C^3 are the second and third conserved cysteines of the EGF repeat, S/T is the modified Ser/Thr residue, and X can be any residue) *(22)*. A second type of O-fucosylation has been identified. On thrombospondin type 1 repeats (TSRs), a disaccharide form of O-fucosyl glycans (Glcβ1–3Fucα1-Ser/Thr) is found on the human extracellular matrix protein "thrombospondin-1" *(24)*. TSRs are found in many extracellular proteins. A single TSR is ~60 amino acids long and is characterized by conserved Cys, Trp, Ser, and Arg residues. The putative consensus sequence site for this modification is $WX_5CX_{2/3}S/TCX_2G$ *(22)*.

O-GLYCAN CONSENSUS SITES

For most O-glycosylation types, a recognition consensus sequence for the attachment of the first sugar residue remains unknown. The exceptions are the O-Glc and O-Fuc modifications, for which putative consensus sites have been described [see above and Refs. *(21, 22)*]. The lack of a consensus sequence can arise from the coexist-