Introduction

We have described a major new form of protein glycosylation (termed O-GlcNAc) that is found in all eukaryotes from yeast to man. O-GlcNAc is distinct from other types of protein-bound saccharides in several respects (for review see [1]): (1) O-GlcNAc is a simple, unmodified monosaccharide moiety glycosidically linked to the side-chain hydroxyls of serine or threonine, often occurring at multiple sites on the same protein. (2) Virtually all O-GlcNAc moiety glycosidically linked to the side-chain hydroxyls of serine or threonine, often occurring at multiple sites on the same protein. (2) Virtually all O-GlcNAc is found within the nucleoplasmic and cytoplasmic compartments of cells. (3) O-GlcNAc appears to be highly dynamic and responsive to cellular stimuli in a fashion analogous to phosphorylation.

In this short review, we will first outline evidence localizing O-GlcNAc to the nucleoplasmic and cytoplasmic cellular compartments and discuss proteins that bear O-GlcNAc. Subsequently, we will illustrate what is known about the sequence localization of O-GlcNAc attachment sites and the enzymes that specifically attach or remove the saccharide. Finally, we will describe our working hypotheses about the functions of this ubiquitous eukaryotic protein modification.

O-GlcNAc is highly localized to nucleoplasmic and cytoplasmic proteins

Studies probing the saccharide topology of murine lymphocytes with purified bovine milk galactosyltransferase [2] surprisingly revealed that the majority of terminal, protein-bound N-acetylgalactosamine residues of lymphocytes are O-glycosidically linked monosaccharide moieties (O-GlcNAc). Even more unexpectedly, detergent latency studies showed that nearly all of this novel but abundant (B lymphocytes have ~1.5 x 10^8 O-GlcNAc/cell) saccharide–protein linkage occurs on intracellular proteins. Recent studies using T-lymphocyte hybridomas and more refined methods of primary lymphocyte isolation, have shown that O-GlcNAc is undetectable on the surfaces of viable lymphocytes, but is highly restricted to their nucleoplasmic and cytoplasmic compartments [3]. Subcellular fractionation studies in rat liver indicate that a large number of different proteins are modified by O-GlcNAc with roughly 55% localized within the nucleus, 35% localized to the cytoplasm, and much less in other organelles [4]. The nuclear envelope is particularly enriched in O-GlcNAc. However, the bulk of the O-GlcNAc is found on chromatin proteins. Several groups have documented that many of the proteins comprising the nuclear pore complex are glycosylated at their cytoplasmic and nucleoplasmic faces by O-GlcNAc [5–8]. Likewise, O-GlcNAc is undetectable on the surfaces of intact human erythrocytes, but is detected in striking abundance when detergent-permeabilized cells are probed with galactosyltransferase [9]. Several intrinsic membrane proteins of the endoplasmic reticulum have been shown to bear cytosolically oriented O-GlcNAc residues [10]. Monoclonal antibodies specific for the GlcNAc residues of streptococcal group-A carbohydrates, which also recognize O-GlcNAc-bearing proteins, bind to the nuclear periphery, nucleoplasm and cytoplasm of mammalian cells, but do not bind to plasma membrane [11]. Consistent with these localization studies, all of the well-characterized proteins identified to date that bear O-GlcNAc are known to reside within either the nucleoplasm, cytoplasm or both (see below). Finally, latency studies on a recently identified UDP-GlcNAc:protein O-GlcNAc glycosyltransferase have directly shown that the enzyme’s active site is localized within the cytoplasmic compartment [12].

O-GlcNAc occurs on a myriad of proteins, most of which are also phosphorylated and multimeric

To date, only a few of the hundreds of proteins that bear O-GlcNAc have been identified (Table 1). The first O-GlcNAc proteins to be unequivocally identified were the nuclear pore proteins [5]. Many different monoclonal antibodies prepared against purified nuclear envelopes recognize various subsets of at least 10 distinct nuclear pore proteins. In nearly every case, O-GlcNAc moieties comprise a major portion of the antibodies’ epitopes [5]. These and
Structure and Function of Glycosylation

Table I

Identified O-GlcNAc-bearing proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human erythrocyte band 4.1</td>
<td>[9]</td>
</tr>
<tr>
<td>Bovine brain synapsins</td>
<td>[19]</td>
</tr>
<tr>
<td>Cytokeratins 13, 8, 18</td>
<td>[29, 30]</td>
</tr>
<tr>
<td>Vertebrate lens α-crystallins</td>
<td>[31]</td>
</tr>
<tr>
<td>Adenovirus fibre proteins</td>
<td>[34, 35]</td>
</tr>
<tr>
<td>65 kDa Nuclear tyrosine phosphatase</td>
<td>[28]</td>
</tr>
<tr>
<td>RNA polymerase II</td>
<td>[26]</td>
</tr>
<tr>
<td>v-Erb-A oncogene protein</td>
<td>[24]</td>
</tr>
<tr>
<td>Human cytomegalovirus phosphobasic tegument protein</td>
<td>[33]</td>
</tr>
<tr>
<td>Nuclear pore complex proteins</td>
<td>[5–13]</td>
</tr>
<tr>
<td>Human erythrocyte 65 kDa soluble protein</td>
<td>[9]</td>
</tr>
<tr>
<td>Eukaryotic initiation factor 2-associated</td>
<td>[18]</td>
</tr>
<tr>
<td>67 kDa protein (p67)</td>
<td></td>
</tr>
<tr>
<td>RNA polymerase II transcription factors</td>
<td>[21–23]</td>
</tr>
</tbody>
</table>

O-GlcNAc is also found on a large number of proteins tightly associated with *Drosophila* chromatin [20]. These glycoproteins are uniformly distributed along the length of polytene chromosomes. A fluorescently tagged, GlcNAc-binding lectin (wheat-germ agglutinin (WGA)) stains the polytene chromosome with a banding pattern that is remarkably similar to that seen for chromosome-binding dyes. Interestingly, the apparent density of O-GlcNAc is substantially reduced at ‘puff’ regions (active sites of transcription) in these polytene chromosomes. Several RNA polymerase II transcription factors, but not factors for RNA polymerase I or III, have been shown to bear O-GlcNAc [21]. Since this initial report, additional transcription factors [22, 23] and oncogene proteins [24] have also been shown to be glycosylated. Interestingly, WGA inhibits the transcriptional activation, but not the DNA-binding by the transcription factor Sp1 [21]. WGA–Sepharose affinity chromatography has proven to be a valuable step in the purification of RNA polymerase II transcription factors [25]. The large catalytic subunit of RNA polymerase II itself is also glycosylated by O-GlcNAc in its C-terminal domain (CTD) [26]. The CTD is a region of the protein containing a heptapeptide sequence (consensus sequence YSPTSPS) that repeats 52 times in the mammalian enzyme [27]. Previous studies have shown that the CTD is essential for cell viability and is rapidly and extensively phosphorylated during its transition from an initiating to an elongating transcription complex. The non-phosphorylated form of the RNA polymerase II contains O-GlcNAc, whereas the phosphorylated form does not contain detectable amounts of the saccharide [26], suggesting that perhaps a reciprocal relationship between phosphorylation and glycosylation of the enzyme may exist. While still inconclusive, these data imply that O-GlcNAc may play a role in regulating the initiation of the rapid phosphorylation cascade that accompanies transcriptional elongation.

While only a very small proportion of the O-GlcNAc proteins have been identified to date, it is already clear that the saccharide is found on proteins with widely diverse structures and functions. In addition to those proteins mentioned above, O-GlcNAc has been found on an interesting 65 kDa nuclear tyrosine phosphatase that reversibly associates with chromosomes at mitosis [28]. Intermediate filament proteins, cytokeratin 13 [29], and cytokeratins 8 and 18 [30] are also glycosylated by O-GlcNAc. Metabolic labelling studies on these cytokeratins indicate that their O-GlcNAc residues...
turnover much more rapidly than their protein backbones. Vertebrate lens \(\alpha\)-crystallins, which are phosphorylated structural proteins homologous to the small heat shock protein hsp27, are the most abundant O-GlcNAc-modified proteins identified so far [31]. The abundance of these proteins has allowed precise localization of the major site of O-GlcNAc attachment. Mass spectrometry (m.s.) of the \(\alpha\)-crystallins has also confirmed that the monosaccharide moieties are not elongated or further modified. Recently, the \(\alpha\)-crystallins have been shown to be authentic heat shock proteins in several non-lens tissues [32]. While only a few studies have examined O-GlcNAc in viruses, it is already clear that many viral proteins will probably be found to be modified by O-GlcNAc. Studies of the human cytomegalovirus (human herpesvirus 5. HCMV) 149 kDa basic phosphoprotein (BPP) demonstrate that it contains O-GlcNAc [33]. The BPP is a tegument protein which accounts for \(\sim 20\%\) of the virion mass. The presence of glycoproteins in the virus tegument is unprecedented. These studies also indicated that several other CMV proteins may also be modified by the saccharide. O-GlcNAc residues are also found on adenovirus fibre proteins, which are responsible for virus attachment to the cell surface [34, 35]. However, in this case, most of the O-GlcNAc residues appear to be buried within the polymerized fibre and are inaccessible to galactosylation. Parasitic protozoa, such as African trypanosomes (W. G. Kelly & G. W. Hart, unpublished work) and Schistosomes [36] synthesize numerous O-GlcNAc bearing proteins. O-GlcNAc represents the major type of O-linkage found in Schistosoma mansoni, a parasitic blood fluke which causes hepatosplenic disease in humans [36].

Thus, even though O-GlcNAc is found on an enormous diversity of proteins, some common features are evident. Virtually all of the O-GlcNAc-modified proteins are also phosphorylated by specific kinases in a regulated manner. Nearly all are intranuclear, or are found within the nucleus in certain cell types or at certain stages of the cell-cycle or under certain metabolic conditions. In addition, the O-GlcNAc proteins reversibly and specifically associate with other proteins to form multimeric complex structures. For example: (a) The nuclear pore proteins dissociate and re-associate with the pore complex at each cell cycle [37, 38]. (b) The synapsins reversibly associate with the cytoplasmic face of synaptic vesicles, depending upon their phosphorylation state (refs. in [19]). (c) Band 4.1 reversibly associates with glycoporphin and spectrin (refs. in [9]). (d) The cytokeratins reversibly form interme-

diate filaments [29, 30]. (e) The p65 tyrosine phosphatase reversibly associates with chromosomes [28]. (f) The eukaryotic initiation factor-2 (eIF-2) associated p67 is thought to promote protein synthesis by reversibly associating with eIF-2 initiation factor preventing its phosphorylation [18, 39]. (g) RNA polymerase II and its associated transcription factors reversibly form multi-component transcriptional complexes that exquisitely control the transcription of specific genes [40].

**O-GlcNAc attachment sites are not precisely determined at the primary sequence level**

Unlike N-linked glycosylation, which almost invariably occurs at -Asn-X-Ser- sequences [41], but analogous to conventional-type O-linked glycosylation [42], the sequences recognized by the O-GlcNAc glycosyltransferase(s) are not obvious at the primary sequence level. Fig. 1 shows the attachment sites that have been identified thus far. Common features of these sites include a proline residue within one to three amino acids from the attachment site. The O-GlcNAc moieties also appear to occur in Ser/Thr-rich regions of the proteins. Studies with bacterial hexosaminidases, as well as with galactosyltransferase labelling of non-denatured proteins, suggest that on most proteins the O-GlcNAc moieties are exposed on the surfaces of the molecules. The small amounts of most O-GlcNAc-modified proteins available, the presence of O-GlcNAc at many sites in sub-stoichiometric amounts, and the difficulties of gas-phase sequencing glycosylated residues in Ser/Thr-rich

**Fig. 1**

Sites of O-GlcNAc addition

Diamonds denote sites of glycosylation. Italic serines and threonines are other potential sites which have not yet been confirmed.

Nuclear pore 62 kDa:

Met-Ala-Gly-Pro-Ala-Asp-Thr-Ser-Asp-Pro-Leu

Red blood cell 65 kDa:

Adp-Ser-Pro-Val-Gln-Pro-Ser-Leu-Val-Gly-Ser-Lys

Red blood cell band 4-1:

Ala-Gln-Thr-Ile-Thr-Ser-Glu-Thr-Pro-Ser-Thr

Bovine \(\alpha\)-crystallin:

Asp-Ile-Pre-Val-Ser-Arg-Glu-Glu-Lys
domains, have greatly impeded localization of O-GlcNAc attachment sites. However, recent advances in m.s. techniques are alleviating these difficulties [43]. Clearly, localization of the specific attachment sites for O-GlcNAc is essential for future site-directed mutagenesis studies designed to examine the functional significance of the saccharide.

Cytosol contains enzymes that attach and remove O-GlcNAc
Based upon the O-GlcNAc attachment sites identified (Fig. 1), we synthesized a peptide (YSDDSPSTST) for use as a substrate in assays designed to identify and purify an O-GlcNAc:peptide glycosyltransferase. Using such assays, we have identified [12] and purified to apparent homogeneity a cytosolic UDP-GlcNAc:peptide N-acetylglucosaminyltransferase from rat liver [44]. Other workers have also shown that translated recombinant p62 nuclear pore protein in vitro is efficiently glycosylated by reticulocyte lysate cell-free translation systems indicating that sufficient UDP-GlcNAc is already present in these lysates [45]. The rat liver O-GlcNAc transferase is freely soluble, has a very high affinity for UDP-GlcNAc ($K_m \sim 500 \mu M$), and lacks a metal-ion requirement. The purified enzyme appears to be a large phosphorylated protein (>320 kDa) comprised of two subunits in a molar ratio of roughly two to one (~110 kDa and 78 kDa, respectively). Photoaffinity probes suggest that the ~110 kDa subunit contains the catalytic site of the enzyme. Systematic analyses of synthetic peptides clearly demonstrate that the enzyme is remarkably specific. The transferase appears to require nearby proline residues, and is strikingly affected by the distance of the hydroxy amino acid from the proline moiety. Current efforts are defining the minimal peptide substrates. Molecular modelling of these peptides will hopefully elucidate common three-dimensional structural features.

The rapid turnover of O-GlcNAc [30, 46] suggests that cytosolic or nuclear N-acetylglucosaminidases may also be important to the saccharide's biological activity. Several laboratories have identified and partially characterized neutral, cytosolic N-acetylglucosaminidases [47]. We have identified and purified to near homogeneity such an enzyme from rat spleen [48]. The rat spleen enzyme's pH optimum is 6.4, and it does not bind to Concanavalin A-Sepharose. The enzyme, unlike lysosomal hexosaminidases, has no activity against GalNAc, and is highly efficient at removing O-GlcNAc from peptides.

The cloning of cDNAs encoding these enzymes and the determination of how their expression and activities are regulated will help to determine the factors influencing the levels of O-GlcNAc on various proteins. In addition, cDNA probes and antibodies will be invaluable in determining if there is a family of O-GlcNAc transferases/glycosidases analogous to the kinases/phosphatases which have been described for phosphorylation systems.

Is O-GlcNAc a regulatory modification analogous to phosphorylation?
O-GlcNAc is on a myriad of different intracellular proteins and appears to be nearly as abundant as phosphorylation. As illustrated in Fig. 2, metabolic labelling studies with $[^{3}H]$glucosamine clearly show that little, if any, of the O-GlcNAc is elongated or modified. Gel filtration analyses in 6 M-guanidine-HCl and 4 M-urea confirmed that all of the protein-associated $[^{3}H]$GlcNAc in these studies was originally covalently attached to the protein. The lack of further modification or elongation of O-GlcNAc is supported by m.s. analyses [23, 31, 43].

Thus, the question remains: Why do all eukaryotic cells attach a neutral monosaccharide such as GlcNAc to specific serine or threonine sites

![Fig. 2: O-GlcNAc is not elongated](image)
on so many different types of proteins? To determine if O-GlcNAc is a dynamic modification responsive to cellular stimulation, we examined its levels in murine lymphocytes and T-cell hybridomas after stimulation by antigens or by phorbol esters and ionophores [46]. Apparent levels of O-GlcNAc on many nuclear proteins increase rapidly after a short time of stimulation (minutes to hours depending upon the stimuli and on the protein examined). Likewise, the apparent levels of O-GlcNAc on many cytosolic proteins rapidly decrease after stimulation. In both cases, the levels appear to return to initial values after a short period of time, suggesting a cyclic pattern. As indicated above, recent radiolabelling studies of O-GlcNAc on cytokeratins [36] also indicate that the saccharide turns over much more rapidly than the protein. Based upon these initial observations, O-GlcNAc does appear to be highly dynamic in a manner quite similar to protein phosphorylation.

In summary, O-GlcNAc is an abundant post-translational modification which may be as widely distributed on intracellular proteins as phosphate. The saccharide occupies sites on proteins that are very similar to sites recognized by several serine/threonine kinases. In addition, the O-GlcNAc attachment sites have a high 'PEST' score [49], suggesting that O-GlcNAc might specifically regulate proteolysis. The common features of the O-GlcNAc-bearing proteins suggest that the saccharide might be involved in mediating important protein-protein associations. Finally, the finding that most O-GlcNAc-bearing proteins are nuclear has led to the suggestion that it might be involved in nuclear targeting [8]. A recent study demonstrated that the phosphorylation of a specific serine residue that is not in the nuclear localization sequence of SV40 T antigen can nonetheless regulate the targeting of that protein [50]. Thus, it remains possible that O-GlcNAc could play a role in the nuclear targeting process even though the peptide signals for nuclear localization do not themselves contain glycosylation sites. Our working hypothesis is that O-GlcNAc is a regulatory modification which has more properties in common with phosphorylation systems than it has with other well-studied types of glycosylation. With the tools now at hand, the firm elucidation of O-GlcNAc's function should soon be forthcoming.

blood group

galactose-containing glycoconjugates

results from the addition of GalNAc

An even greater enhancement of binding activity

rabbit antiserum raised against 3T3 cell-derived

nucleoprotein; hnKNP, heterogenous nuclear KNP;

snKNI', small nuclear RNP particles; Mac-2, macrophage

cell-surface antigen.

Introduction

Carbohydrate-binding protein 35 (CBP35; \( M_r \sim 35,000 \)) was initially purified from extracts of mouse 3T3 fibroblasts on the basis of its binding to galactose-containing glycoconjugates [1]. The affinity of the lectin for the disaccharide lactose is 60-100 times greater than that for galactose. Moreover, the binding affinity is further enhanced \( \sim 10 \) -fold by the incorporation of an acetamido group at position 2 of Glc (i.e. to yield N-acetyllactosamine). An even greater enhancement of binding activity results from the addition of GalNAc \( \alpha \) 1 at position 3 of the Gal moiety of lactose derivatives (e.g., as in blood group A tetrasaccharide; see [2, 3]).

Immunoblotting analysis, using a polyclonal rabbit antiserum raised against 3T3 cell-derived CBP35, showed that the lectin and its homologues in other species are found in a variety of tissue types and various cell lines [2, 3]. The amino acid sequence (264 residues), deduced from the nucleotide sequence of a cDNA clone [4], showed that the CBP35 polypeptide consisted of two domains (Fig. 1a): an N-terminal half that contains repeats of the sequence Pro-Gly-Val-Pro-Gly followed by three other amino acids (Pro-Gly-rich domain) and a C-terminal half that is homologous to other galactose-binding proteins classified as S-type lectins by Drickamer [5]. Lectins are grouped into S-type families on the basis of conserved amino acid residues within a characteristic carbohydrate recognition domain (CRD), which is clearly distinguishable from the corresponding CRD of the calcium-dependent C-type lectins. Within a 39-residue sequence, 15 amino acid residues are invariant in the CRDs of all S-type lectins (Fig. 1a). The S-type lectins are divided in turn into two subgroups: L-14 and L-30. In general, the polypeptide of the L-14 group consists of a single domain, the CRD (\( M_r 10,000-16,000 \)). The L-30 group poly-