A family of UDP-GalNAc: polypeptide N-acetylgalactosaminyl-transferases control the initiation of mucin-type O-linked glycosylation

Henrik Clausen\(^1\) and Eric P.Bennett

Department of Oral Diagnostics, School of Dentistry, Faculty of Health Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark

\(^1\)To whom correspondence should be addressed at Department of Oral Diagnostics, School of Dentistry, Faculty of Health Sciences, University of Copenhagen, Norre Alle 20, Copenhagen N, DK-2200, Denmark

Key words: glycosyltransferases/mucins/O-glycosylation/GalNAc-transferases

Enzymatic glycosylation of proteins involves the addition of a monosaccharide or, in the case of N-linked glycosylation of asparagine, a preformed oligosaccharide to an amino acid in a given protein. The initial step of protein glycosylation is an important event in the formation of a given glycopeptide linkage (glycoconjugate type), which involves essential recognition events between the protein and glycosyltransferase that determine the specific sites of glycan attachment. Complete processing and extension of glycan chains involves the cooperative action of perhaps hundreds of different glycosyltransferases, which successively add monosaccharides to the growing glycan chain. The characterization of glycan structures on glycoproteins as well as the identification of specific sites of glycan attachment are important for understanding the structure of a given glycoprotein, its function, and its immunobiology (Lis et al., 1993; Varki, 1993; Parekh, 1994).

Mucin-type O-glycosylation is initiated by the enzyme UDP-\(N\)-acetylgalactosamine: polypeptide \(N\)-acetylgalactosaminyltransferase (GalNAc-transferase) (EC. 2.4.1.41). In this reaction GalNAc (derived from the donor substrate UDP-GalNAc) is transferred to the side chains of serine and threonine residues on polypeptides. This reaction, which has been named mucin-type O-glycosylation because of its preponderance on mucin-like glycoproteins, is one of the most abundant forms and means of glycosylation found in animals and is not restricted to mucin-like glycoproteins. This review will cover recent developments in our understanding of a family of GalNAc-transferases that mediate this type of protein glycosylation.

Mucin-type glycosylation is an important post-translational modification of many animal proteins. Alterations in the O-glycosylation patterns of some proteins may play a role in the pathogenesis of several diseases in which there is evidence for altered glycoprotein structure, including ulcerative colitis, chronic bronchitis, cystic fibrosis, and cancer (Varki, 1993). In addition to its structural and functional importance to mucins, in which O-linked carbohydrates may constitute up to 80% of the total mass of these glycoproteins, O-glycosylation has been shown to be important to the folding of proteins such as hCGβ, and the conformation and protease resistance of ‘stem regions’ of membrane-bound proteins. Mucin-like domains are also found on cell surface associated molecules functioning as se-lectin ligands (e.g., PSGL-1, MadCAM, CD-34; see Varki, 1993).

A long-standing puzzle has been what determines sites of O-glycosylation. This is an important question because of evidence that spacing of O-glycans on many glycoproteins may be important to their function (Springer, 1994). The molecular processes governing the specificity and kinetics of mucin-type O-linked glycosylation and the parallel Man-type glycosylation of serine/threonine in yeast are poorly understood. It has not been possible thus far to determine precise positions of glycan attachment to the protein backbone for many glycoproteins. This is due, in part, to the difficulties of addressing these questions experimentally because mucins and mucin-like glycoproteins are often insensitive to protease digestion and glycans are often clustered in arrays, making sequence analysis difficult (Piller and Piller, 1993; Gooley et al., 1994).

Consensus motifs for glycosylation have been described for prediction of several types of glycosylation sites other than mucin-type O-glycosylation. The best characterised example is that of N-linked glycosylation of asparagine where the consensus sequence for the acceptor site has been resolved to a short peptide sequence Asn-Xaa-Ser/Thr (where Xaa cannot be Pro) (Marshall, 1972), although Asn-Xaa-Cys has also been found to be utilized. Recent studies indicate that Asn-Xaa-Ser is less well utilized compared to the Thr containing sequon (Kasturi et al., 1995). Studies of other types of glycopeptide linkages also suggest the existence of rather specific primary peptide sequences associated with glycosylation. Thus, proteoglycan-type glycosylation of serine is restricted to -Ser-Gly-Xaa-Gly- (Bourdon et al., 1987), and collagen-type glycosylation of hydroxylysine restricted to -Gly-Xaa-Hyl-Gly- (Procop et al., 1979). The GlcNAc-type glycosylation of serine or threonine appears to be adjacent to an acidic amino acid and within two residues of a proline (Haltiwanger et al., 1992). The Fuc-type glycosylation of serine/threonine seems to be restricted to the peptide sequence -Gly-Gly-Thr/Ser-Cys- (Harris and Spellman, 1993), and recently a prokaryotic O-linked glycan type was found associated with Asp-Ser- (Plummer et al., 1995).

In contrast, a defined peptide sequence for mucin-type GalNAc O-glycosylation and the yeast parallel Man-type O-glycosylation of serine/threonine are not evident despite considerable efforts to identify such a motif (Sadler, 1984; Gooley et al., 1991, O’Connell et al., 1991; Wilson et al., 1991; Elhammer et al., 1993; Gooley and Williams, 1994). A number of studies have attempted (unsuccessfully) to identify a consensus sequence for animal GalNAc O-glycosylation by evaluating similarity among sequences around known sites of glycosylation (Sadler, 1984; Gooley et al., 1991, O’Connell et al., 1991; Wilson et al., 1991; Elhammer et al., 1993; Gooley and Williams, 1994), and by testing the peptide substrate specificity of the GalNAc-transferase activity in crude and pure form (Gooley et al., 1991; O’Connell et al., 1992; O’Connell and Tabak, 1993; Wang et al., 1993; Nishimori et al., 1994a,b).
The results of those studies can be explained by either of two hypotheses: there is a single human GalNAc-transferase with broad acceptor substrate specificity or there are several GalNAc-transferases with different substrate specificities. Here, we discuss recent data that support the latter of these hypotheses. These studies establish that a family of polypeptide GalNAc-transferases are expressed in animal cells. These enzymes have different acceptor substrate specificities and are differentially expressed in different cells and organs. Similarly, it has recently been shown that mannosyl O-glycosylation in yeast, which is mediated by enzymes adding mannose to serine and threonine sites of protein via dolichyl phosphate-d-mannose, are mediated by a family of Man-transferases with distinct acceptor substrate specificities (Strahl-Bolsinger et al., 1993; Lussier et al., 1995). Insight into the specificity and role of individual members of this GalNAc-transferase family of enzymes may significantly aid in understanding the complexity of O-glycosylation and reveal more defined primary sequence motifs with predictive value of O-glycosylation positions.

**Purification of polypeptide GalNAc-transferases have resulted in identification of two distinct enzymes**

Polypeptide GalNAc-transferases transfer N-acetylgalactosamine from a sugar-nucleotide donor (UDP-GalNAc) to serine or threonine residues (McGuire et al., 1967; Roseman, 1970). GalNAc-transferase activity was originally partially purified and characterized from submandibular glands (Hagopian and Eylar, 1969; Hill et al., 1977). The enzyme activity requires divalent metal ions as a cofactor, with Mn2+ being optimal (Sugiura et al., 1982; Elhammer and Kornfeld, 1986). Enzyme activity is optimal over a broad pH range between 6.5 and 8.6 (Hagopian and Eylar, 1969; Elhammer and Kornfeld, 1986). The $K_m$ for the donor substrate UDP-GalNAc is approx. 8 μM (Elhammer and Kornfeld, 1986) and $K_m$ for acceptor substrates using peptides varies from 25 μM with a 30-mer peptide derived from the Mucl tandem repeat being one of the best substrates identified thus far (Nishimori et al., 1994b).

Early reports of purification of GalNAc-transferase to apparent homogeneity employed an affinity chromatography step using deglycosylated submaxillary mucin (apomucin) as a ligand for binding of GalNAc-transferase in the presence of UDP (Sugiura et al., 1982; Elhammer and Kornfeld, 1986). Homa et al. (1993) used this procedure to isolate colostrum GalNAc-transferase from which partial amino acid sequence information was obtained. In subsequent studies O’Connell et al. (1992) and Hagen et al. (1993) isolated the same bovine colostrum GalNAc-transferase using Hg-UDP-GalNAc affinity chromatography, which was initially developed by Bendik and Schachter (1987). Wang et al. (1992) also utilized the Hg-UDP-GalNAc affinity chromatography for the purification of a porcine submandibular GalNAc-transferase, and recent partial protein sequencing data suggested that the isolated porcine enzyme was identical to the bovine GalNAc-transferase (Roth et al., 1994). The soluble bovine colostrum enzyme was found to be an N-linked glycoprotein of approximately 70,000 (Elhammer and Kornfeld, 1986), which is similar in apparent size to that found for the purified porcine submaxillary gland (Wang et al., 1992). N-Glycanase treatment reduced the size of the bovine enzyme to approximately 64,000 (Homa et al., 1993) and the porcine gland enzyme to approximately 60,000 (Wang et al., 1992).

A different GalNAc-transferase has been isolated from human placenta using a peptide affinity chromatography strategy that was modified from the apomucin chromatography (White et al., 1995). This strategy was employed because of the possibility that affinity chromatography matrices with multiple acceptor substrate sites or the use of nucleotide ligands can result in the purification of multiple types of GalNAc-transferases. The use of a different purification strategy resulted in the isolation of a novel GalNAc-transferase distinct from the one isolated previously (White et al., 1995). The two isolated enzymes are encoded by distinct genes (see later section) and designated GalNAc-T1 (bovine colostrum enzyme) (Homa et al., 1993; Hagen et al., 1993) and GalNAc-T2 (human placenta) (White et al., 1995).

Sugui et al. (1982) originally purified a soluble GalNAc-transferase of approximately 55,000 (as ascertained by SDS-PAGE and gel filtration analysis) from rat ascites hepatoma, AH66. The recently isolated human placental GalNAc-transferase migrated according to a molecular weight of approximately 52,000 (White et al., 1995). Because of difficulties in purifying glycosyltransferases, the successful isolation of glycosyltransferases has generally only been fully acknowledged after protein sequencing and subsequent gene cloning and expression have conclusively demonstrated functional activity. In this respect the early study of Sugui (Sugui et al., 1982) has been partly ignored because of a discrepancy in molecular weight as compared to the bovine colostrum enzyme that was subsequently purified. Importantly, the predicted mass of the soluble form of GalNAc-T2 is slightly higher than that of GalNAc-T1 but apparently the two apoproteins migrate in SDS-PAGE and gel filtration quite differently. Thus, GalNAc-T1 migrates as a 64,000 protein after N-glycanase treatment (Homa et al., 1993), whereas GalNAc-T2 likely not to be N-glycosylated migrates as a 52,000 protein; yet the predicted mass of both is around 59,000. It is likely that the rat enzyme purified by Sugui (Sugui et al., 1982) may have represented the human placental GalNAc-transferase, i.e., a GalNAc-T2 rat analogue.

**A family of polypeptide GalNAc-transferases controls O-glycosylation**

The GalNAc-T1 transferase purified from bovine colostrum was cloned independently by Homa et al. (1993) as well as Tabak and colleagues (O’Connell et al., 1992; Hagen et al., 1993). This enzyme was found to have a human equivalent with a highly similar nucleotide sequence that had 79 nucleotide substitutions encoding only 6 amino acid differences (Meurer et al., 1995; White et al., 1995). Recently, the porcine and rat GalNAc-T1 analogues have been sequenced and found to be virtually identical to the bovine and human versions indicating that this gene is highly conserved (Hagen et al., 1995; Yoshida et al., 1995). In this discussion we have used the human GalNAc-T1 sequence as reference for alignment purposes. The human placental GalNAc-T2 transferase purified and cloned by White et al. (1995) shares only 45% amino acid sequence similarity with GalNAc-T1.

As shown in Figure 1A, a potential GalNAc-transferase motif conserved within family members has been identified in a 61 amino acid stretch, with 80–85% amino acid sequence similarity (Bennett et al., 1996). A similar motif was identified for the sialyltransferase family, and motifs have been identified for other glycosyltransferases. The existence of these highly conserved sequences suggests that these regions are functionally important or are directly involved in the common acceptor or donor substrate sites. The major motifs shared between si-
A family of GalNAc-transferases

Polypeptide GalNAc-Transferase Motif

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>GalNAc-T1</td>
<td>GGENLEISFR1WQCGTLEIVTCSHVGVERKATPYTFPGGTGQIINQNRRALAEVMDE</td>
</tr>
<tr>
<td>GalNAc-T2</td>
<td><em><strong><strong><strong><strong><em>V<em><strong><strong>S</strong></strong></em>IP</em>RRRRRRRQH</strong></strong><em><strong><em>S</em>TVFAR+TA</strong></em></strong></strong></em></td>
</tr>
<tr>
<td>GalNAc-T3</td>
<td><em><strong><strong><em><strong><strong>IY<em><strong><strong>V</strong></strong></em>Q</strong></strong>MP</em>V</strong></strong></em>SKS<em>HS</em>K*----V<em>AR</em>QA*-----</td>
</tr>
<tr>
<td>GalNAc-T4</td>
<td><strong><strong><strong><strong><strong>K</strong></strong></strong></strong></strong>HPI***<em><strong><strong><em><em>IP</em>RA</em>AR<em>N-------FLO</em>TA*A</strong></strong></em></td>
</tr>
<tr>
<td>ZK688.8</td>
<td><em><strong><strong><strong><strong><em>V</em>M</strong></strong>*S</strong></strong></em>HP<em>RRRRRRRQ********AKV</em>HH<em>AAT</em>-----</td>
</tr>
</tbody>
</table>

B

Plot similarity of ClustalW Multiple Sequence Alignment of GalNAc-T1, -T2, -T3, and -T4.

Fig. 1. (A) Amino acid sequences found in the GalNAc-transferase motif. Multiple sequence alignment analysis (ClustalW) reveal amino acid similarities of 80-82% in this region in contrast to the overall similarity of 40-50%. GenBank/EMBL accession numbers of genes analysed are as follows: human GalNAc-T1, #X85018; human GalNAc-T2, #X85019; human GalNAc-T3, #X92689; and C.elegans, #L16621. Note that the N-terminus of the C.elegans encoded protein cannot be inferred from the DNA sequence submitted to GenBank, but this has been identified in an exon 5' to the submitted sequence (Hagen et al., 1995). (B) Comparison of amino acid sequences of human GalNAc-T1, -T2, -T3, and -T4 analysed by plot similarity of ClustalW multiple sequence analysis. Sequence similarities are restricted to the central part of the transferases. Position of the GalNAc-transferase motif used for the isolation of GalNAc-T3 and -T4 is indicated (Bennett et al., 1996). Positions of aligned cysteine residues are indicated by vertical arrows (see Fig. 2).
GalNAc-transferases: Conserved cysteine residues

A

Fig. 2. (A) Primary structure of human GalNAc-transferases T1-T4 and the C. elegans ZK688.8 homologue. Multiple protein sequence alignment analysis (ClustalW) shows that the central 80% of the sequences align nearly perfectly, and 12 cysteine residues throughout this region are conserved in spacing (indicated by lines). The differences in sequence lengths of the human GalNAc-transferases are attributed to variable sequences in the region immediately C-terminal of the putative transmembrane regions. The transmembrane regions are indicated by white boxes with numbers of flanking residues. The positions of the N-terminal ends of the purified soluble bovine GalNAc-T1 (Homa et al., 1993) and human GalNAc-T2 (White et al., 1995) are indicated by arrows including residue numbers. The N-terminal end of a potential soluble GalNAc-T3 has not been determined; the arrow shows the N-terminal end of a functionally active protein expressed in the baculo-virus system (Bennett et al., 1996). GalNAc-T4 has not been expressed as yet. Indicated by lollipops are potential N-glycosylation sites, and the position of the GalNAc-transferase motif used for PCR cloning of GalNAc-T3 and -T4 is indicated by a line.

B

Hopp and Woods Hydrophobicity Plot of GalNAc-Transferases

Although most studies agree that O-glycosylation is initiated in the cis-Golgi compartment (Roth, 1984; Deschuyteneer et al., 1988; Piller et al., 1990; Pascale et al., 1992; Wilson et al., 1993), several studies have suggested that ER and intermediate ER-Golgi compartments may be involved (Tooze et al., 1988; Perez-Vilar et al., 1991; Crommie and Rosen, 1995). It is unclear, whether all the individual members of the GalNAc-transferase family localize in the same subcellular location. Recent studies indicate that at least some of the GalNAc-transferase activities identified so far are located in the Golgi apparatus and not in ER or the ER-Golgi intermediate compartment (Schweizer et al., 1994). In an immuno-EM study GalNAc-T1 was localized to cis-Golgi (Roth et al., 1994), but the extent to which the anti-serum utilized for these studies may cross-react with other GalNAc-transferases has not been demonstrated. Furthermore, immunolocalization studies of tagged GalNAc-T1, -T2, and T3 cDNA constructs expressed in HeLa cells have shown that all these localize to the Golgi apparatus (Rötger et al., 1996).

The exact border between the stem region and the catalytic
domain cannot be inferred from available data. However, GalNAc-T1 has a stem region of at least 13 amino acids and T2 at least 26 amino acids as estimated by the N-terminal sequence obtained from the purified soluble enzymes (Homa et al., 1993; White et al., 1995). The differences in total length of GalNAc-transferases appear to be located in the stem region as shown in Figure 2. Henion et al. (1994) have characterized the C-terminal border of the stem region of an α1-3galactosyltransferase and shown that a total of 67 amino acids C-terminal of the identified N-terminus of the soluble active transferase could be removed without destroying the catalytic activity. Recently, Boegegeman et al. (1995) showed that the N-terminal 129 residues of the β1-4galactosyltransferase were unimportant for the kinetic properties of the transferase. These findings suggest that the N-terminus of the soluble, catalytically active enzymes still retain a major part of peptide sequence that is not required for activity and hence may be considered to be part of the stem region. This fact together with the complete lack of sequence similarities between the GalNAc-transferases in the N-terminal region lead to the prediction that size differences in GalNAc-transferases result in differences in the length of the stem regions. It is hypothesized that one function of the stem region is to allow the catalytic domain to protrude into the Golgi lumen for better access to the protein substrates (Paulson and Colley, 1989). The functional effect of the apparent differences in length of the stem region in GalNAc-transferases is unknown. One attractive possibility is that this domain of the protein is involved in determining substrate specificity of the enzymes, e.g., types of acceptor substrate proteins. This may not be true for O-glycosylation of mucins, which has repetitive hydrophilic sequences. Other protein substrates may exhibit conformational restrictions for access of the GalNAc-transferases, such as those that receive glycosylation of a single isolated site in proteins with complex tertiary structure.

All mammalian GalNAc-transferases identified to date contain potential N-linked glycosylation sites (Figure 2). The purified GalNAc-T1 was found to be an N-linked glycoprotein with two to three out of four potential N-linked glycosylation sites utilized (Homa et al., 1993). However, GalNAc-T2 only includes one potential N-linked glycosylation site at the very C-terminal end of the protein, and this site was apparently not or at least only partly utilized in the human enzyme as evidenced by peptide sequencing (White et al., 1995). The consensus site in GalNAc-T2 is -Asn-Asp-Ser-, and recent studies have demonstrated that serine may be incompletely utilized for N-glycosylation (Kasturi et al., 1995). Recently, we have found that GalNAc-T2 expressed in the baculo-system is indeed not glycosylated, as evidenced by matrix-assisted laser desorption time-of-flight mass spectrometry of the purified protein (K.Mirgorodskaya, P.Roepstorff, and H.Clausen, unpublished observations); GalNAc-T3 encodes four and GalNAc-T4 two potential N-linked glycosylation sites.

One striking similarity among the GalNAc-transferases is the spacing of cysteine residues. No less than 12 cysteine residues can be aligned between GalNAc-T1, -T2, -T3, and -T4, each containing a total of 16, 13, 16, and 15 cysteines, respectively (Figure 2). Recently, Hagen et al. (1994, 1995) have identified a genomic sequence found in C.elegans (GenBank/EMBL accession #L16621) with homology to the polypeptide GalNAc-transferase genes. The coding region of this gene predicts a protein with overall amino acid sequence similarity of 40-50% compared to the human GalNAc-transferases. The identified C.elegans gene may encode a polypeptide GalNAc-transferase as indicated by expression studies (F.K.Hagen and L.A.Tabak, personal communication). The putative C.elegans protein designated ZK688.8 shows a slightly higher degree of similarity to GalNAc-T1 than to -T2, -T3, or -T4, but the similarity is restricted to the central region found to be similar within all GalNAc-transferases, whereas the N-terminal and C-terminal are completely dissimilar. In the C.elegans protein sequence 10 of the cysteines are also aligned with the human GalNAc-transferases, whereas the other two cysteines aligned in GalNAc-T1, -T2, -T3, and T4 (sequence: -Cys-Glu-Cys-) are substituted by valine residues (sequence: -Val-Glu-Val-) (Figure 2). Drickamer (1993) originally noted three conserved cysteine residues in the sialyltransferase family. Intramolecular disulphide bridging as well as involvement of cysteine residues in substrate binding or catalysis has been observed in the β1-4-galactosyltransferase (Yadav and Brew, 1991; Wang et al., 1994; Boegegeman et al., 1995). Holmes et al. (1995) have noted a conserved pattern of cysteine residues in five members of the fucosyltransferase family. In this work they were involved in intermolecular disulfide bridging, which is important for in vitro activity. Given the high number of conserved cysteine residues in the GalNAc-transferase family it is likely that these are important and involved in intramolecular disulfide bonding and/or substrate binding and catalysis. Thus, the rather low sequence similarity between the GalNAc-transferase members, it is conceivable that they retain a similar overall three-dimensional structure in agreement with their related substrate specificities. Interestingly, it has been observed that GalNAc-transferase activities are enhanced two- to threefold by the addition of reducing agents to the reaction mixture (Wang et al., 1992; White et al., 1995).

A number of expressed sequence tag (EST) clones encoding potential homologous sequences have been identified in C.elegans, suggesting that C.elegans also contains a family of polypeptide GalNAc-transferases, and some of these have been partially characterized (Hagen et al., 1995; F.K.Hagen and L.A. Tabak, personal communication). It will be interesting to evaluate the relationship of individual human GalNAc-transferases with their C.elegans counterpart in terms of conservation of protein sequence and enzymatic function.

Differential organ expression of GalNAc-transferases

Polypeptide GalNAc-transferase activity has been measured in a variety of organ extracts as well as in a number of cell lines including fibroblasts. To the best of our knowledge, no animal cell type has been found to be devoid of GalNAc-transferase activity. Careful analyses of GalNAc-transferase activities in various organs reported in the literature may reveal discrepancies in their acceptor substrate specificity. These discrepancies may be related in part to experimental variations as discussed later; however, it is likely that such differences also reflect different levels of individual GalNAc-transferases with varying acceptor substrate specificity. Analysis of the acceptor substrate specificities of enzyme extracts from different organs has revealed distinct differences (Hagen et al., 1995; Sørensen et al., 1995).
The expression of individual GalNAc-transferases appears to be variable in different cells and organs. Northern blot analysis showed that GalNAc-T1 and GalNAc-T2 are more or less constitutively expressed in most organs; however, especially GalNAc-T1 shows some quantitative variation (Homa et al., 1993; White et al., 1995; Bennett et al., 1996). In contrast, GalNAc-T3 showed a variable expression pattern in which highest expression was detected in pancreas and testis, weak expression was detected in a few other organs, and no expression was detected in brain, liver, muscle, thymus, or fetal organs (Bennett et al., 1996) (Figure 3). It is important to note that these Northern blot experiments were performed with commercial human multiple mRNA blots, which are based on entire organ extracts. Thus, the results reflect a total picture of expression in several different cell types as well as cell differentiation stages, and only immunohistological or in situ hybridization studies may provide deeper insight into the cell-specific expression patterns.

The full sequence of mRNA transcripts for any of the GalNAc-transferases is not known. GalNAc-T1 hybridizes to two mRNAs of approximately 3.4 and 4.1 kb, and the difference in these transcripts is suggested to be a result of multiple polyadenylation signal rather than differences in 5' untranslated sequence (Meurer et al., 1995). GalNAc-T2 to one mRNA of 4.5 kb, and GalNAc-T3 to a 3.6 kb mRNA. GalNAc-T3 may be differentially spliced or may use alternative promoters, as a larger transcript of approximately 4.2 kb was detected in spleen (Bennett et al., 1996).

GalNAc-transferase genes and their chromosomal localization

A number of glycosyltransferase genes have been cloned and characterized (Joziasse, 1992; Kleene and Berger, 1993; Schachter, 1994). The coding region of glycosyltransferases is found in single or multiple exons. The number of coding exons has varied between one and seven where the largest number of coding exons was found in the histo-blood group ABO gene (Bennett et al., 1995a). Preliminary studies of the genomic organization of the human GalNAc-transferases have revealed that GalNAc-T1, -T2, and -T3 are encoded by an even larger number of exons ranging between 10 and 16 (Bennett et al., 1995b; Meurer et al., 1996). In striking contrast, the coding region of GalNAc-T4 resides in a single exon (Figure 4) (E.P. Bennett and H.Clausen, unpublished data). The coding region of the C.elegans gene is contained in at least six exons. No information is available on the regulatory elements of these GalNAc-transferase genes, but considering that the individual GalNAc-transferases are differentially expressed in organs future efforts should be undertaken in this direction.

The number and positions of intron/exon boundaries within

![Fig. 3. Northern blot analysis of human tissues. Multiple human Northern blots (MTNI, II and fetal) from Clontech as labeled was probed with 32P-labeled GalNAc-T1 probe (TEB1), GalNAc-T2 probe (TEB2), and GalNAc-T3 probe (TEB3), respectively. Reproduced with permission from Bennett et al., 1996](https://academic.oup.com/glycob/article-abstract/6/6/635/635426).
A family of GalNAc-transferases

Fig. 4. Genomic structure of the coding regions of human GalNAc-transferase and the C. elegans genes. The genomic structure of the C. elegans gene is reported with the GenBank/EMBL submission Human GalNAc-T1 structure is deduced from a PI clone containing the entire coding sequence, human GalNAc-T2 structure is deduced from multiple PI clones combined with PCR mapping of one intron, and human GalNAc-T3 and -T4 deduced from single PI clones (E.P.Bennett and H.Clausen, unpublished observations). The 5' ends of the most 5' exons and the 3' ends of the most 3' exons (untranslated sequences) have not been determined, and these are indicated by grey boxing.

the GalNAc-transferase genes support their evolutionary relationship and suggest an order of appearance. GalNAc-T1 and -T3 genes appear to share at least six intron/exon boundaries and four of these are also shared with GalNAc-T2 (Figure 4). The boundary shared between all these genes in the 5' end nearly divides the non-similar most 5' sequences with the similar 3' sequences. The boundary shared in the 3' end is apparently also found in the C. elegans gene. Remarkably, GalNAc-T4 is encoded by a single exon and thus does not contain the shared intron boundaries suggesting that GalNAc-T4 is likely to represent an earlier gene in evolutionary terms than GalNAc-T1, -T2, and -T3, as well as the C. elegans gene. Although the GalNAc-T4 gene contains an open reading frame encoding a predicted 578 amino acid protein, we have not yet been able to express and demonstrate a functional transferase.

Recently, Meurer et al. (1996) identified a GalNAc-transferase pseudogene exhibiting 78% sequence similarity to GalNAc-T1, but, with several nonsense and frame shift mutations. This pseudogene apparently lacked exon 2 when compared to GalNAc-T1 organisation but only had two intronic sequences which were positioned differently from introns in the GalNAc-T1 gene.

Chromosomal localizations of four human GalNAc-transferase genes have been determined by fluorescence in situ hybridization to metaphase chromosomes using PI phages (Figure 5) (E.P.Bennett, A.G.van Kessel, and H.Clausen, unpublished observation). GalNAc-T1 localized to chromosome 18q12-21, GalNAc-T2 to 1q4.1-4.2, GalNAc-T3 to 2q24-31, and GalNAc-T4 to 12q21.3-q22, showing that the GalNAc-transferase family does not co-localize to a single locus. Meurer et al. (1995) have also localized GalNAc-T1 to chromosome 18 by somatic cell hybrids, and GalNAc-T1 related pseudogene to chromosome 3. In three previously described glycosyltransferase families two or several genes have been found to colocalize. The human blood group ABO gene and a truncated version of the α1-3-galactosyltransferase gene co-localize to 9q33–34 (Shaper et al., 1992; Bennett et al., 1995a), several members of the fucosyltransferase family have been co-localized to chromosome 19 (Weston et al., 1992), and two members of the β6-GlcNAc-transferase family have been co-localized to chromosome 9q21 (Bierhuizen et al., 1995).

Do the acceptor substrate specificities of polypeptide GalNAc-transferases determine positions of O-glycosylation?

O-Glycosylation is thus performed by a family of GalNAc-transferases encoded by distinct genes and differentially expressed in cells. The need for this redundancy appears to be very old as evidenced by the C. elegans studies as well as studies of mannosyl-O-glycosylation in yeast. An obvious question is then what functional significance this has for O-glycosylation. Our favoured hypothesis is that multiplicity of GalNAc-transferase is required for O-glycosylation of different peptide sequences; however, the available data supporting this hypothesis are ambiguous.

In vivo studies. The analysis of in vivo O-glycosylation by structural characterization of isolated glycoproteins is one direct approach to identify potential preferred sequences for O-glycosylation which can be used to infer information about the acceptor substrate specificity of GalNAc-transferases. This approach, however, suffers from several experimental shortcomings. The available analytical techniques to identify glycosylation sites are inadequate. Moreover, the presently available data are derived from selected glycoproteins and do not include mucin-type glycoproteins, which are the most heavily glycosylated proteins in this class (Gooley et al., 1991; O’Connell et al., 1991; Wilson et al., 1991; Elhammer et al., 1993). The studied glycoproteins have been isolated from a variety of organs and species, and O-glycosylation specificity may vary in different cells. Data related to this point has been reviewed by Gooley and Williams in this journal (Gooley and Williams, 1994), to which we refer. The main conclusions from these studies are that no specific peptide sequences appear to determine O-glycosylation, but that Pro, Ser, and Thr are favorable in adjacent positions, whereas charged and large aromatic amino acids may interfere with glycosylation.
or tertiary structure which allowed O-glycosylation, and that around the existing O-glycosylation site contained secondary substitutions in the near vicinity of the only naturally O-glycosylated tertiary charges at position -1 reduced O-glycosylation.

The overall results indicate that in vivo O-glycosylation of this site is not significantly influenced by alterations in flanking sequences. The major problem in attempting to reconcile previous in vitro results with the in vivo results is that the former analysis was performed with a single enzyme, while the in vivo analysis includes the activity of multiple GalNAc-transferases. It is likely that the ability of cells to glycosylate different substrates is at least in part due to the activities of several GalNAc-transferases in those cells.

In another recent study, the in vivo O-glycosylation of recombinant erythropoietin was studied in detail (Elliott et al., 1994). Elegant analyses of 33 independent serine and threonine substitutions throughout the molecule showed that only substitutions in the near vicinity of the only naturally O-glycosylated serine (a.a. 126; -ISPPDAAS\textsuperscript{126}AAPL- ) were glycosylated in the substituted proteins. The authors concluded that the region around the existing O-glycosylation site contained secondary or tertiary structure which allowed O-glycosylation, and that the primary amino acid sequence by itself was not important. It is possible that a completely opposite interpretation of these data is correct. The two mutations leading to de novo glycosylation in vivo were (glycosylated residue is underlined): -ISPPT\textsuperscript{123}AASA-; -PPDAT\textsuperscript{125}SAAP-; and substitution of Ser\textsuperscript{126} with Thr was also glycosylated -PDAAT\textsuperscript{126}AAPL-.

In the same region, the peptide having the sequence -SPPDT\textsuperscript{124}ASAA- was not glycosylated at the threonine, which in fact would be expected from previous in vitro studies of specificity (O'Connell et al., 1992), because of the negatively charged Asp in position -1. In contrast, all other substitutions that did not lead to O-glycosylation were positioned in primary peptide sequences which would not be predicted to serve as substrates by the method of Elhammer et al. (1993). It is our opinion that these data can also be taken as evidence supporting primary sequence requirement for O-glycosylation. A challenging experiment would be to correlate these in vivo findings with in vitro assays of O-glycosylation using appropriate GalNAc-transferases and peptides covering the 33 substitutions.

In vitro studies. In the past, a number of studies have attempted to define the acceptor substrate specificity of GalNAc-transferase activity using glycosylation of apoproteins, fragments thereof, and more recently synthetic peptides (Sadler, 1984; Wilson et al., 1991; O'Connell et al., 1991; Gooley and Williams, 1994). It is difficult to interpret the results obtained from these studies in light of findings supporting the existence of an entire family of polypeptide GalNAc-transferases. GalNAc-transferase activities in total organ extracts, semipurified preparations, or even in apparently homogenous GalNAc-transferase preparations may reflect the combined activities of several or different transferases. For example, differences in acceptor substrate specificities have been demonstrated at different steps of the purification of human placental GalNAc-transferase (Sørensen et al., 1995). Moreover, in vitro GalNAc-transferase analysis may not take into account the role of potential co-factors which could influence the kinetics and substrate specificities (Brodbeck et al., 1967; Do et al., 1995). Finally, potentially proteolytic digestion of acceptor substrates, interaction of blocking peptides, or competing GalNAc-transferases (O'Connell et al., 1992; Sørensen et al., 1995), may create artifacts that affect the interpretation of the results of these studies.

Wang et al. (1992) have proposed that GalNAc-transferases show no acceptor specificity when analysed with synthetic peptides lacking the conformational constraints of proteins. Gooley and Williams (Gooley and Williams, 1994) have suggested that in vitro studies may be less useful, and instead stressed the importance of pursuing studies of in vivo O-glycosylation pathways in order to achieve insight into the regulation of O-glycosylation. Although this is a worthwhile goal, it is difficult to perform for every potential substrate. Therefore, we advocate determination of substrate specificities of the GalNAc-transferases in studies using the individual recombinant GalNAc-transferases to complement the in vivo studies. Despite drawbacks, several studies have now shown that primary peptide sequences influence GalNAc-transferase activity in assays; therefore, such studies should yield information about the recognition domain of GalNAc-transferases and the influence of acceptor substrate sequence on enzyme activity.

The most comprehensive study demonstrating this thus far was performed by O'Connell et al. (1992), who analyzed the specificity of purified bovine colostrum GalNAc-transferase...
(later shown to represent GalNAc-T1) on a set of 12-mer peptides derived from the sequence around the known in vivo O-glycosylation site of human von Willebrand factor. Systematic substitutions of each amino acid around the central Thr acceptor site with Ile, Ala, Pro, and Glu generated 53 peptides, which were analysed as substrates for the purified bovine trans-
ferase. The results clearly show that subtle changes in the primary structure in all positions -6 to +5 of the Thr influenced the in vitro glycosylation markedly. These studies did not provide evidence supporting any obvious general rules for O-linked glycosylation; nevertheless, substrate specificity observed in test tube assays was in good agreement with what has been learnt from identified glycoprotein, notably, a clear posi-
tive effect of Pro substitutions in positions +1, +2, and +4, and a negative effect of charged residues in position -1. However, as discussed above, these results are not in agreement with a recent study of in vivo O-glycosylation in COS cells (Nehrke et al., 1996).

Studies on in vitro glycosylation of the Muc1 20-mer tandem repeat (-PDTRPAPGSTPAPPAGVTS-) show that the two threonines in -TS- and -ST- are utilized and to a lesser degree the S in -ST-, whereas the isolated threonine in -DTR- is not glycosylated (Nishimori et al., 1994a,b; Stadie et al., 1995). Importantly, the length of the peptides analyzed did not have a major influence on glycosylation of this substrate since identical glycosylation sites were observed with short 9-mer pep-
tides (Nishimori et al., 1994a), 20-mer peptides covering 1 tandem repeat, as well as 60-mer or 105-mer peptides covering 3 and 5 tandem repeats, respectively (Nishimori et al., 1994b; H.Hassan, K.Mirgorodskaya, P.Roepstorff, and H.Clausen unpublished observations). The kinetic parameters, in terms of \(k_m\) and \(V_{max}\), however, are better for longer peptides, and this is a general phenomenon observed for all peptide designs (Nishimori et al., 1994b). Fontenot et al. (1994) have demonstrated that the 60-mer Muc1 repeat peptide retains a tertiary structure which the 20-mer peptide does not; however, in vitro O-glycosylation is identical for the two peptide designs (Nishimori et al., 1994b). Stadie et al. (1995) have demonstrated that glycosylation using extracts from T47D cells showed an ordered sequence of GalNAc incorporation into the Muc1 sites with T in -TS- being the first site saturated and T in -ST- being the second. This finding may suggest that a special repertoire of GalNAc-transferases is expressed in T47D cells, because analysis of GalNAc-transferase activities in organ extracts showed that some extracts more readily glycosylated a partial tandem repeat peptide containing -TS-, whereas others showed better activity with a peptide containing the -ST- sequons (Sørensen et al., 1995). Preliminary analysis with recombinant GalNAc-T1 to -T3 has revealed similar clear differences in kinetics; however, the end result after prolonged reaction time (24-48 h) for glycosylation showed the same three sites utili-

Analysis of 32 overlapping 15-mer peptides covering the entire sequence of the HIV envelope glycoprotein gp120 showed that a limited number of peptides were acceptors for human placenta or porcine submandibular gland GalNAc-
transferase preparations (Clausen et al., 1994). Out of 85 poten-
tial Ser/Thr sites in the 32 peptides, 5 peptides containing a single or multiple Ser/Thr residues were found to be acceptors for GalNAc-transferase activity. The best substrate site was identified in the V3 loop of gp120 (-GPGRAVTGKNNM-), which represents an unlikely acceptor substrate based on past in vivo O-glycosylation information (Elhammer et al., 1993), yet later studies have shown that a specific GalNAc-transferase (GalNAc-T3) selectively exhibits this activity (Bennett et al., 1996). Although the gp120 in vivo O-glycosylation sites remain unknown, O-glycosylation of this protein is reported to be 4-5 mol per mole (Merkle et al., 1991; Hansen et al., 1992).

As pointed out earlier, most of the in vitro studies so far have been performed with GalNAc-transferase preparations which may contain multiple GalNAc-transferases even when purified to apparent homogeneity. Emerging availability of recombinant GalNAc-transferases will aid in understanding the sub-
strate specificities. Homa et al. (1993) have expressed a full length of bovine GalNAc-transferase in the baculo-virus sys-
tem, but detailed studies of the specificity of this enzyme have not yet been presented (Meurer et al., 1995). Hagen et al. (1993) have expressed the soluble bovine GalNAc-transferase and provided preliminary characterization of its specificity (Hagen et al., 1994). We have expressed the soluble form of the human GalNAc-transferase designated GalNAc-T1, the hu-
man placenta GalNAc-transferase designated GalNAc-T2, and the recently cloned GalNAc-T3 transferase (Bennett et al., 1996). Preliminary analysis of the substrate specificities of these recombinant enzymes clearly show that they display both distinct and overlapping acceptor specificities toward a panel of synthetic peptides. Substrates exclusively utilized by one or the other enzyme have been identified showing that the en-
zymes do exhibit distinct specificities (Bennett et al., 1996). Clearly, detailed studies with large panels of peptide substrates are needed before more general conclusions can be drawn.

It is clear from the above data that in vitro O-glycosylation of peptides is far from random and that it is influenced by the primary amino acid sequence of the acceptor sites. The extent to which the overall protein structure participates in determin-
ing sites and rates of O-glycosylation is still quite unclear. It remains important to determine if in vitro O-glycosylation correla-
tes with in vivo O-glycosylation in order to determine if peptides are useful for illustrating potential consensus motifs for the individual GalNAc-transferases.

Conclusions and future perspective

The first and key event of mucin-type O-glycosylation is the initia-
tion step where GalNAc is added to selected serine and
threonine residues of proteins. Recent evidence clearly indi-
cates that this step is considerably more complex than previ-
ously believed. It is now clear that an entire family of poly-
peptide GalNAc-transferases with at least three and possibly
more members are involved in this process. The 'central dogma of glycobiology'—one (glycosidic) linkage, one en-
zyme—as originally proposed by Hagopian and Eylar (1968) was insightful and has been a guideline for decades for iden-
tification and characterization of specific glycosyltransferase

A family of GalNAc-transferases
activities. As predicted, at least one specific glycosyltransferase has been identified for each known linkage; perhaps unexpectedly, it is now clear that for several linkages an entire group of enzymes are involved. These display subtle differences in substrate specificity and their expression is differentially regulated in different organs. This level of complex regulation suggests that a high level of functional significance accompanies these particular glycosidic linkages.

Mucin-type O-glycosylation occurs in all animal cells. The data discussed in this review show that patterns of O-glycosylation in terms of attachment sites of glycans may vary from cell to cell depending on the expressed repertoire of the polypeptide GalNAc-transferases. This adds a new level of complexity to the study of protein glycosylation where the past focus has been on the diversity of further processing of the glycans.

The polypeptide GalNAc-transferase family is extremely old in evolutionary terms, as evidenced by the existence of multiple members in C.elegans. The apparent need for multiple enzymes with different acceptor substrate specificities strongly indicates that each of the enzymes is functionally important for proper O-glycosylation. Future studies should analyze the acceptor substrate specificity of each of the enzymes using in vitro studies as well as by using in vivo model systems with gene knock-outs or de novo introduction of specific GalNAc-transferases. It has long been anti-dogmatic to believe that GalNAc-transferase activity in cells should show selectivity for acceptor substrates because of the diversity of peptide sequences found in vivo identified O-glycosylation sites. However, even the likelihood that multiple GalNAc-transferases are responsible for O-glycosylation in a cell, it is conceivable that more restricted peptide sequences can be characterized as acceptor sites for individual polypeptide GalNAc-transferases.

We envision that changes in expression of GalNAc-transferases in various pathological disorders including cancer may result in aberrant glycosylation, e.g., exposure of naked peptide regions on mucins normally covered by glycans. Furthermore, because of redundancy in functions among family members, it appears that experimentally selective knock-out of a specific GalNAc-transferase may not lead to dramatic effects per se. In fact, a gene ablation of a GalNAc-transferase displaying 93% amino acid identity to the GalNAc-T1 appears to result in no phenotypic changes (Hennet et al., 1995). Hence, the existence of null alleles for any particular GalNAc-transferase gene may be asymptomatic or give rise to subtle changes in biological behaviour.

Acknowledgements

We are grateful to Drs. Danielle Weghuis and Ad Geurts van Kessel for performing the FISH experiments, and Drs. Anne Kristensen, Kajsa Mortensen, Oddskaya, and Peter Roepstorff for performing the structural analysis. We thank Drs. Fred K.Hagen and Lawrence A.Tabak for communicating preliminary results prior to publication. We are grateful to Michael A.Hollingsworth for extensive advice and suggestions for the manuscript. Studies performed in our laboratory were funded by the Danish Medical Research Council, Danish Natural Science Research Council, The Lundbeck Foundation, The Danish Cancer Society, Velux Foundation, Dr. Ib Henriksen’s Foundation, Gangsted Foundation, Novo Nordisk Foundation, Carlsberg Foundation, and Ingeborg Roijer’s Foundation.

References


Received on January 4, 1996; revised on May 16, 1996; accepted on May 20, 1996.