Glycoconjugate Journal 16, 171–180 (1999)
© 1999 Kluwer Academic Publishers. Manufactured in The Netherlands

The acceptor specificity of UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases

Åke P. Elhammer*, Ferenc J Kézdy, and Akira Kurosaka[1]

Pharmacia & Upjohn, 301 Henrietta St, Kalamazoo, MI 49007;[1] current address: Kyoto Sangyo University, Division of Biotechnology, Faculty of Engineering, Kamigamo-Motoyama, Kita-Ku, Kyoto 603-8555, Japan

The *in vitro* and *in vivo* specificity of the family of peptide:N-acetylgalactosaminyltransferases (GalNAcT) is analyzed on the basis of the reactivity and/or inhibitory activity of peptides and protein segments. The transferases appear to be multi-substrate enzymes with extended active sites containing a least nine subsites that interact cooperatively with a linear segment of at least nine amino acid residues on the acceptor polypeptide. Functional acceptor sites are located on the surface of the protein and extended conformations (β-strand conformation) are preferred. The acceptor specificity of GalNAc-T can be predicted from the primary structure of the acceptor peptide with an accuracy of 70 to 80%. The same GalNAc-T enzymes catalyze the glycosylation of both serine and threonine residues. The higher *in vitro* catalytic efficiency toward threonine *versus* serine is the result of enhanced binding as well as increased reaction velocity, both effects being the result of steric interactions between the active site of the enzyme and the methyl group of threonine. Results from substrate binding studies suggest that GalNAc-T catalyzed transfer proceeds via an ordered sequential mechanism.

Keywords: **O-glycosylation, UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases, substrate specificity**

Introduction

Mucin-type O-linked oligosaccharide structures are common constituents of many secretory and cell surface proteins [1,2]. They are characterized by the N-acetylgalactosamine which invariably links the oligosaccharide chain to the polypeptide backbone by means of an O-linkage to a serine or threonine residue in the protein [3]. Mucin-type O-linked structures may occur either as single entities at one distinct site on polypeptides or, perhaps more frequently, as clusters attached to specific sequence segments [e.g., 4–16]. Such sequence segments, first identified in mucins and commonly referred to as "mucin domains," are rich in serine, threonine, and proline, while they contain relatively few charged or strongly hydrophobic amino acids and few strong determinants of secondary structure [15–20]. Generally, mucin domains consist of repeating amino acid sequence segments each of which may contain a number of glycosylated serine and threonine residues [15–17]. Within a given mucin repeat segment, the saccharide structures can be located as close to each other as being attached to contiguous amino acids [4,21,22], frequently resulting in molecules where the saccharide structures contribute more than half of the molecular mass.

Initiation of mucin type O-linked oligosaccharide biosynthesis

The biosynthesis of mucin-type O-linked oligosaccharides is initiated by the transfer of an N-acetylgalactosamine from UDP-GalNAc to a serine or threonine residue on an acceptor polypeptide. This reaction is catalyzed by a family of enzymes collectively referred to as UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferases (GalNAc-T). At the time of writing this review, several distinct GalNAc-T sequences have been identified in mammals [23] and expression of five of these, namely GalNAc-T1, GalNAc-T2, GalNAc-T3, GalNAc-T4, and GalNAc-T5, has been described [24–29]. The sequences of the five enzymes all encode active GalNAc-transferases with what appears to be distinct but largely overlapping substrate specificities [28–30]. While the sequence similarity between these five GalNAc-transferases is rather limited, it is nevertheless significant and indicates common origin and common functionality. As shown in Figure 1, in the sequence comparison of nine enzymes, the segments corresponding to positions 115–156 and 285–375 in the human GalNAc-T1 display a selectively high degree of identity, which sug-

*To whom correspondence should be addressed: Åke P. Elhammer, 7240-209-622, Pharmacia & Upjohn, 301 Henrietta St., Kalamazoo, MI 49001-4940; E-mail: ake.elhammer@am.pnu.com

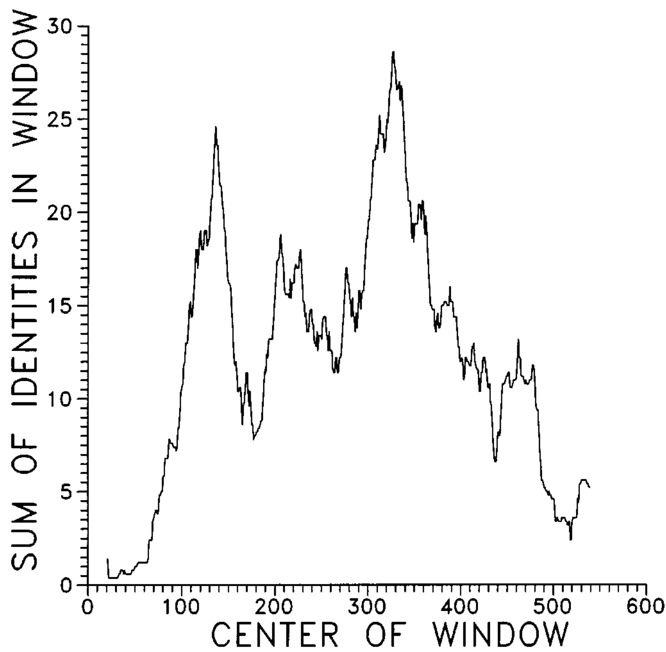

Figure 1. The amino acid sequences of human GalNAc-T1, bovine GalNAc-T1, rat GalNAc-T1, mouse GalNAc-T1, human GalNAc-T2, human GalNAc-T3, mouse GalNAc-T3, mouse GalNAc-T4, and *C. elegans* GLY3CE were aligned using the CLUSTAL algorithm as first described by Clausen and Bennett [30]. The resulting identity graph—with the sequence positions numbered according to the human GalNAc-T1—was then scanned using a window of 41 amino acids within which the number of identities was summed, with conservative replacements counting as 0.2 identity. The two peaks showing $<50\%$ identity, i.e., $Sum_{41} \geq 21$, correspond to segments 115–156 and 285–375.

gests that they constitute part of the active site. The segment 316–376 shows 80–85% sequence similarity. This segment has been designated as the "GalNAc-T motif" and is proposed to be important for substrate binding [31].

In contrast to N-linked sugars, O-linked oligosaccharides are conjugated post translationally to the acceptor proteins: Available data suggest that the transfer of the anchoring N-acetylgalactosamine occurs either in an intermediate compartment between the endoplasmic reticulum and the Golgi apparatus [32–34] or in the *cis* portion of the Golgi stack [35–41]. However, only limited information is available concerning the transfer reaction *in situ.* A UDP-GalNAc transport system, analogous to other nucleotide sugar antiporters identified in the secretory pathway, has been shown to be present in a highly enriched rat Golgi membrane preparation [42–43]. This system may then be responsible for transporting the UDP-GalNAc required for the reaction from the site of its synthesis in the cytoplasm, to the luminal compartment where the glycosylation is presumed to take place. Other details of the *in vivo* transfer of N-acetylgalactosamine, such as the local topology of the glycosylation machinery, the format of acceptor presentation within the glycosylation compartment and the actual, *in situ,* conditions for the reactions, remain unknown. Moreover, the individual, specific roles of the several GalNAc-T enzymes identified to date have not yet been recognized. Although the five mammalian GalNAc-transferases expressed to date appear to have somewhat different *in vitro* acceptor specificities, those differences are essentially quantitative since most acceptors tested are glycosylated by all five enzymes, albeit with differing catalytic efficiencies [28–30]. Also, Nehrke et al. [44], while studying the *in vivo* acceptor specificity of GalNAc-T, found that a reporter molecule containing a single glycosylation site derived from the human von Willebrand factor was equally glycosylated—with $>95\%$ efficiency—by Cos 7, L6 and by 10(3) cells, even though GalNAc-T1, -T2, -T3, and -T4 are expressed at quite different levels in the three cell lines.

Catalytic properties of GalNAc-transferases

Specificity toward the donor substrate

GalNAc-T-catalyzed glycosylations have been studied quite extensively *in vitro.* Like many other glycosyltransferases, the GalNAc-transferases have a narrow specificity for their donor substrate, UDP-GalNAc. For GalNAc-T1 this specificity appears to be essentially absolute; other nucleotide sugars have an affinity almost three orders of magnitude lower than that of UDP-GalNAc (Fig. 2). Moreover, the reaction product UDP competitively inhibits the transfer reaction with a K_i which is identical within experimental error to the K_m for UDP-GalNAc (1.7 *vs.* 1.5 μM), suggesting that the major force responsible for the binding of the nucleotide sugar to the enzyme is contributed by the nucleotide portion of UDP-GalNAc and that the interaction with the sugar moiety is essentially a matter of steric fit. The enzyme GalNAc-T3 has a donor substrate specificity similar to that of GalNAc-T1, while another transferase, GalNAc-T2, has a somewhat broader specificity in that it is capable of also using UDP-Gal (30). GalNAc-T1 binds to columns with immobilized *donor* substrate in the presence of Mn^{++}, even in the absence of acceptor [45]. In contrast, the enzyme binds to columns with immobilized *acceptor* substrate only in the presence of the donor substrate, or UDP. This suggests that GalNAc-T1 catalyzed glycosylation reactions proceeds via an ordered sequential kinetic pathway, analogous to that observed with the xylosyl-transferase enzyme which initiates proteoglycan synthesis [46].

Specificity toward the acceptor substrate

The specificity of GalNAc-transferases toward the acceptor substrate is less transparent. Early studies, using crude enzyme preparations or whole cell lysates, as well as more recent investigations using purified or recombinant GalNAc-transferases either intact or in truncated forms, have shown that GalNAc-T is able to transfer N-acetylgalac-

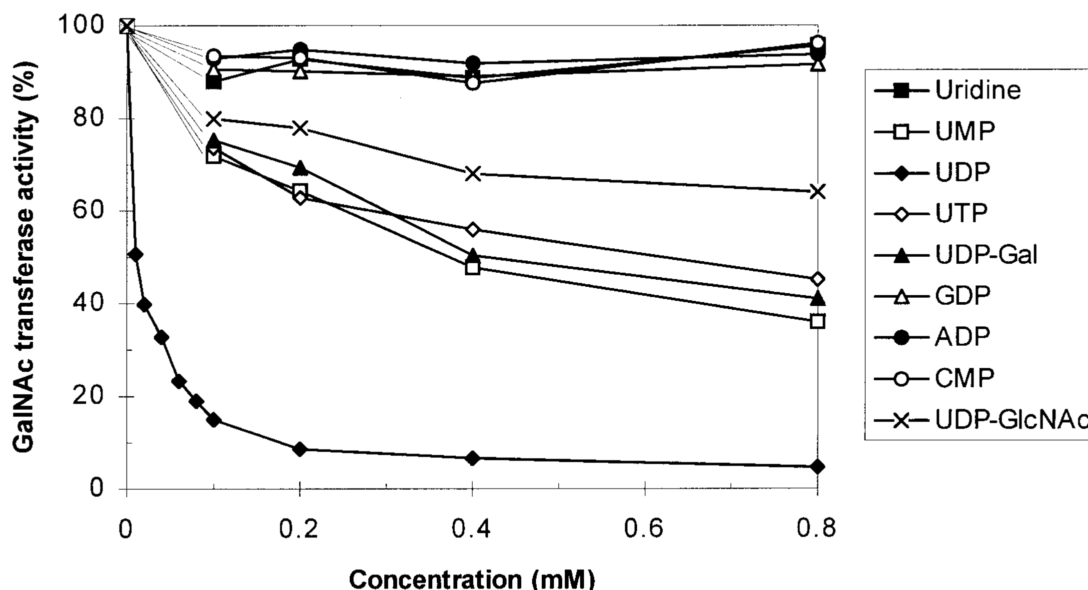

Figure 2. The donor substrate specificity of GalNAc-T1 was investigated by studying the effect of increasing concentrations of uridine and various nucleotides and nucelotide sugars on the activity of the enzyme. Standard assay conditions were employed [18].

tosamine not only to intact and to partially degraded proteins, but also to a number of synthetic peptides of various lengths and diverse compositions [18,47–63]. The influence of polypeptide length on the catalytic activity of GalNAc-transferase was first investigated by Janis Young and co-workers [50,51]. They demonstrated the necessity of the α-amino group on the acceptor amino acid to be blocked for acceptors to be functional, and observed that proline residues located at certain positions toward the carboxy terminal with respect to the acceptor amino acid, can have a dramatic influence on the catalytic activity. Other investigators have confirmed these findings and also found that charged amino acids and/or amino acids with bulky side chains at positions close to the acceptor have a deleterious effect on acceptor efficiency [53–55,57–59,63,64]. Presumably because of their bulkiness, glycosylated amino acids close to the acceptor amino acid also reduce the glycosylation efficiency [62].

A considerable amount of work has been devoted to trying to discover a specific amino acid sequence or a characteristic motif—analogous to the Asn-X-Ser/Thr motif recognized by the oligosaccharyltransferase catalyzing the attachment of N-linked oligosaccharides—that would define the acceptor sites for GalNAc-transferase [65]. Attempts to identify such "motifs" have been made through studies of amino acid sequences surrounding glycosylated serines and threonines in native glycoproteins and through *in vitro* glycosylation studies using proteins and peptides as acceptors [63,64,66–69]. Little, if any, consensus has emerged from these studies. Invariably, many glycosylated serine and threonine residues are surrounded by residues not corresponding to any of the consensus motifs, while the *in vitro* work showed that GalNAc-T1 is capable of glycosylating serines and threonines surrounded by a bewildering variety of sequences. On the other hand, many of the *in vitro* studies do show that even though GalNAc-T's have a wide *in vitro* acceptor specificity, they still display considerable differences in their catalytic efficiency toward different acceptors [18,50,51,53–55,58,59]. Both the composition and sequence of at least four amino acids located at either side of the acceptor have a profound influence on the efficiency of the glycosyl transfer. Also, distribution of certain amino acids at specific positions with respect to the acceptor serine or threonine appears to be associated with high acceptor efficiency [18,54,55,58,59].

In an attempt to delineate the acceptor specificity of GalNAc-T, Elhammer et al. [18] analyzed the amino acid sequences surrounding 196 serine and threonine residues glycosylated *in vivo*. Although no specific sequence(s) or acceptor motifs emerged, the analysis revealed clear non-random distributions of the individual amino acids at sites encompassing up to four residues at either side of the glycosylated amino acids. Serine, threonine and proline, for instance, were significantly surabundant at all nine subsites while essentially all strongly hydrophobic and strongly hydrophilic amino acids occurred at very low frequencies. Still, no evidence was found for an absolute requirement for any given amino acid at any subsite, besides the obvious need for an acceptor serine or threonine. These results indicated that GalNAc-transferase has an extended binding site for the acceptor substrate and that the interaction of the enzyme with its acceptors is most appropriately described by a cumulative subsite mechanism, where the interactions of all the subsites with their corresponding

amino acids in the acceptor sequence independently contribute to the catalytic efficiency, but where none of the interactions (again with the exception of the acceptor amino acid) is an absolute requirement. An extended binding site interaction with the acceptor is also convincingly shown by the importance of sequences upstream and downstream of the acceptor amino acid: several *in vitro* studies demonstrate a clear loss of transfer efficiency upon truncation of the acceptor sequence [50,51,58,59,63]. Moreover, in an *in vitro* study of the glycosylation of human chorionic gonadotropin, Sugahara et al. [70] found that the acceptor site for the glycosylating enzyme(s) extends 4–5 amino acids in each direction from the acceptor serines. They also found, by mutational analysis, that the type of amino acid(s) within the acceptor sequence, with the exception of proline, is not critical for glycosylation and that each (tandem) repeat domain in the carboxy-terminal portion of the protein is glycosylated independently of the state of glycosylation of the other domains.

The specificity of enzymes utilizing a cumulative subsite mechanism can be characterized by a numerical analysis of the frequencies of individual amino acids at each subsite in the acceptor sequence. A set of specificity parameters, defined for each amino acid at each subsite, allows one to calculate the probability that any given amino acid sequence would be recognized by the enzyme. Using such specificity parameters, Elhammer et al. [18] developed an algorithm capable of predicting probable sites for glycosylation by GalNAc-T. The accuracy of the method is about 75–85% [19,21,71,72]. In support of the validity of the predictive algorithm, it was discovered in the course of the analysis of specificity by the cumulative subsite mechanism, that while proteins synthesized in the secretory pathway that are not glycosylated contain very few or no predicted sites for O-glycosylation, other proteins synthesized either in the eukaryotic cytosol or by prokaryotes, may contain a number of potential sites. The latter sites are of course not utilized *in vivo* but can be glycosylated *in vitro.* Although some of these predicted sites are cryptic, i.e., they are not available for glycosylation *in vivo* or *in vitro*—presumably because they are buried in the interior of the protein—they are still functional acceptors for GalNAc-T in that they can be glycosylated *in vitro* upon destruction of the tertiary structure of the protein, for example by reduction and carboxymethylation of the cysteines. Thus, one concludes that the probable sites predicted by this method are indeed functional glycosylation sites and that the specificity parameters determined in the analysis of *in vivo* glycosylated sequences outlined above provide a valid description of the specificity of the enzyme. The results also illustrate that sites for O-glycosylation must be exposed on the surface of the acceptor in order to be glycosylated *in vivo,* a fact consistent with the post-translational localization of this reaction. The majority of predicted sites on mammalian cell surface and secreted proteins are glycosylated *in vivo, i.e.,* in general these proteins contain few unutilized or cryptic sites. Moreover, proteins that are unglycosylated in the native state often contain few if any predicted sites [18], suggesting that proteins not destined for glycosylation *in vivo* may in fact have been selected against the inclusion of glycosylation sites in their sequence.

The specificity parameters obtained in the above analysis allow one to design efficient synthetic peptide acceptors and inhibitors by selecting for each subsite amino acids with the highest specificity. In this way, one can find acceptors that are not segments of native proteins or otherwise derived from native sequences, and yet have a high predicted probability of glycosylation. One such acceptor is the peptide PPAS**T**SAPG, the design of which represents a direct application of the glycosylation specificity parameters discussed above (see ref. [18]). Evaluation of this peptide in *in vitro* glycosylation experiments revealed that it is indeed an efficient acceptor for GalNAc-T-catalyzed glycosylation. In fact, with a catalytic efficiency of 1,330 M^{-1} sec^{-1}, it readily outperforms most peptides based on native acceptor sequences (53,57). One acceptor peptide, PPDAA**T**AAPL, has been found with a higher *in vitro* catalytic efficiency, but its sequence represents a modification of the native sequence, PPDAA**S**AAPL, from human erythropoietin, where threonine has been substituted for the serine in the central, acceptor position. While essentially all GalNAc-T acceptors identified to date have acceptor K_m's in the low mM range, the peptide PPDAA**T**AAPL has a K_m almost an order of magnitude less (Tables 1 to 4) when assayed with GalNAc-T1, but not with GalNAc-T2 or GalNAc-T3.

The acceptor specificity of the GalNAc transferases has also been analyzed by others. Algorithms have been developed that to a modest extent may increase the success rate of predicting potential glycosylation sites. This increase, however, comes at the price of introducing a large number of additional parameters [19,71,72]. The methods are largely heuristic and provide little additional insight into the mechanism of acceptor specificity. It is important to note that all of these analyses are based on the implicit postulates that the glycosyl transferases have an extended acceptor binding site, that the acceptor specificities are the result of the cooperative enhancements contributed by all subsites, and that the same enzyme is responsible for the glycosylation of both serines and threonines.

The analysis by Chou et al. [71] also implies that there may be an interaction between the subsite occupied by the acceptor amino acid and the other subsites, thereby predicting that optimal glycosylation of serines may require the presence of a set of amino acids at the nonacceptor subsites which may be different from that required for the glycosylation of threonines. Since the increase in predictive ability achieved by implying an interaction between the acceptor subsite and the other subsites is statistically insignificant, the interaction, if it occurs at all, should be rather

minor. A strong interaction between the different subsites is also implied by the "multiple motif" specificity model [63,66–69] which posits that when a given subsite is occupied by a certain amino acid, then glycosylation can only occur when the other subsites are occupied by certain corresponding amino acids. In other words, occupancy at a given subsite alters the specificity of the other subsites. The glycosylated protein data base so far failed to reveal any statistically significant cross-correlation between frequencies of occurrence of given amino acids at any two subsites, thereby suggesting that interaction between the different subsites likely is minor, if not negligible.

The amino acid sequences in the immediate neighborhood of the acceptor threonine or serine do not indicate the existence of *subclasses* of acceptor substrate specificities discernible by correlational analyses. And yet, there exists a clear dichotomy of proteins glycosylated by the GalNAc transferases: one class of protein substrates contains domains with multiple repeats of acceptor segments, while proteins in the other class bear only isolated acceptor segments, often consisting of not more than one glycosylable amino acid residue. Glycosylation of proteins in the first class is largely random and often only partial but still yielding fully functional derivatives, while the function of the proteins in the second class depends critically on their glycosylation. Since the reactivity of these two substrate classes is not distinguished by the amino acids bracketing the acceptor threonine/serine function, it is tempting to speculate that in the case of the single-site acceptor proteins, the GalNAc transferases may recognize not only the acceptor enneapeptide segment, but also some structural elements distal to it. Future studies will perhaps shed some light on the existence of these secondary binding sites and the important role that protein conformation may play in the regulation of O-glycosylation.

Relative catalytic efficiency toward serine and threonine-containing acceptors.

Several investigators have reported that GalNAc-transferase glycosylates serine-containing acceptors less efficiently than threonine-containing ones [18,53,55–57]. On the basis of such observations it has even been suggested that GalNAc-T1 is, in reality, a threonine-specific enzyme and that there may exist separate transferases specific for threonine and for serine [53,55]. A comparison of the rates of glycosylation of the threonine and serine containing derivatives of two acceptor sequences, PPAS**T**SAPG/ PPAS**S**SAPG and PPDAA**T**AAPL/PPDAA**S**AAPL, i.e., of substrate molecules that are identical except for the acceptor amino acid, reveals a definite selectivity for threonine, with *in vitro* catalytic efficiencies some 30 to 130 times higher for threonine than for serine (Tables 1 to 4). Nevertheless, all three GalNAc-T enzymes studied are clearly capable of efficiently glycosylating serine residues and the relative efficiencies toward serine and threonine remain the same during all stages of the enzyme purification. Furthermore, the catalytic efficiencies of the serine-containing forms of the above two peptides are comparable to that of some other threonine-containing acceptor peptides whose sequences had been derived from native acceptor "motifs" [69; Table 5]. Also, compilation of all glycosylated sites reported in the literature reveals that the overall number of glycosylated serines and threonines is about equal, suggesting that the two amino acids are equally efficient acceptors *in vivo* [18,19]. These results suggest that the differences between the catalytic efficiencies toward serine and threonine containing acceptors ob-

Table 1. Kinetic parameters of peptide glycosylations catalyzed by GalNAc-T1

Acceptor peptide	K_m *mM*	V_{max} x 10^3 s^{-1}	V_{max}/K_m $M^{-1}s^{-1}$
PPAS-**T**-SAPG	6.9 ± 0.2	9,181 ± 210	1,330
PPAS-allo**T**-SAPG	≈50	≈40	≈0.8
PPAS-**S**-SAPG	7.2 ± 1.6	109 ± 21	15
PPDAA-**T**-AAPL	0.2 ± 0.04	4,901 ± 192	24,505
PPDAA-allo**T**-AAPL	≈17	≈100	≈6
PPDAA-**S**-AAPL	1.8 ± 0.5	435 ± 87	247

Table 2. Kinetic parameters of peptide glycosylations catalyzed by GalNAc-T2

Acceptor peptide	K_m *mM*	V_{max} x 10^3 s^{-1}	V_{max}/K_m $M^{-1}s^{-1}$
PPAS-**T**-SAPG	3.2 ± 0.4	2,456 ± 139	765
PPAS-**S**-SAPG	13.4 ± 0.8	195 ± 18	15
PPDAA-**T**-AAPL	3.2 ± 0.4	2,105 ± 200	658
PPDAA-**S**-AAPL	14.6 ± 0.7	328 ± 10	22

Table 3. Kinetic parameters of peptide glycosylations catalyzed by GalNAc-T3

Acceptor peptide	K_m *mM*	V_{max} x 10^3 s^{-1}	V_{max}/K_m $M^{-1}s^{-1}$
PPAS-**T**-SAPG	3.0 ± 0.3	1,994 ± 214	671
PPAS-**S**-SAPG	14.7 ± 2.7	77 ± 5	5
PPDAA-**T**-AAPL	1.2 ± 0.1	840 ± 51	677
PPDAA-**S**-AAPL	40.2 ± 11.1	489 ± 31	12

Table 4. Catalytic efficiencies, V_{max}/K_m, and Threonine *versus* Serine acceptor specificity ratios

	GalNAc-T1		*GalNAc-T2*		*GalNAc-T3*	
Acceptor peptide	V_{max}/K_m $M^{-1}s^{-1}$	*Thr/Ser*	V_{max}/K_m $M^{-1}s^{-1}$	*Thr/Ser*	V_{max}/K_m $M^{-1}s^{-1}$	*Thr/Ser*
PPAS-**T**-SAPG	1,330		765		671	
		89		51		134
PPAS-**S**-SAPG	15		15		5	
PPDAA-**T**-AAPL	24,505		658		677	
		99		30		56
PPDAA-**S**-AAPL	247		22		12	

served *in vitro,* may be of little consequence *in vivo.* In fact, when expressing an engineered acceptor in COS7 and MCF-7 cells, Nehrke et al. [73] found poor correlation between residues efficiently glycosylated *in vivo* and *in vitro.*

An obvious reason for the efficient glycosylation of serines *in vivo* could be the presence of an as yet undetected serine-specific transferase different from the threonine-specific enzyme. Though still not impossible, the existence of an enzyme that glycosylates exclusively serine residues is somewhat unlikely in view of the fact that in spite of considerable efforts no evidence, direct or indirect, was ever found to point toward the occurrence of such an enzyme [57]. It certainly is not *necessary* to invoke the existence of such an enzyme, since GalNAc-T1, as discussed above, is capable of efficiently glycosylating serine residues *in vitro,* albeit at a slower rate than threonine. Also, the two other GalNAc-transferases tested, GalNAc-T2 and GalNAc-T3, both show the same relative, but not absolute preference as GalNAc-T1, for threonine containing acceptors (Tables 2 and 3). Along these lines, it also has been reported recently that the mammalian GalNAc-T4, and nine GalNAc-transferases identified in *Caenorhabditis elegans,* all glycosylate threonine *in vitro* at a much higher rate than serine, but not exclusively [74].

Against this background it appears likely that enzymes with *in vitro* relative serine/threonine specificities similar to those of GalNAc-T1, -T2, and -T3 are responsible for the bulk synthesis of both glycosylated threonines and serines in the cell. Perhaps, the acceptor concentrations in the *cis*-Golgi compartment where glycosylation takes place, are so high as to make the *in vitro* differences in catalytic efficiencies irrelevant. Alternatively, the lower acceptor efficiency of serine may be compensated for by a longer transit time in the glycosylating compartment. Several mucins, for example, require more than an hour to reach the cell surface and a large portion of this time appears to be spent in transit between the ER and the *trans* Golgi, i.e., in the very compartment(s) where O-glycosylation probably takes place [75–78]. To illustrate the subtle complexity of the question of *in vivo* specificity, consider the fact that Delorme et al. [79] and Elliott et al. [80] observed that—in close analogy with the observations *in vitro*—the threonine residue at position 126 of recombinant erythropoietin expressed in COS1 and CHO cells was glycosylated more efficiently than a serine replacing it. In contrast, Nehrke et al. [73] found that upon expression in COS7 and MCF-7 cells, a recombinant reporter molecule containing one site for O-glycosylation, was glycosylated with equal efficiency whether the site was occupied by a threonine or serine residue. Possible cause for this difference in behavior could be a structural alteration by the serine in erythropoietin, but not in the reporter molecule. Alternatively, the explanation may be as simple as differing transit times in the glycosylating compartment(s). In fact, Nehrke et al. [73] noted that *in vivo* glycosylation of low-efficiency sites could be improved by culturing the cells at a lower temperature, thereby increasing the dwelling time of the reporter molecule in the glycosylating compartment(s). Another possibility is that *in vivo* acceptors are folded and presented to the enzyme in such a way that the efficiency of serine glycosylation is selectively enhanced.

Table 5. Kinetic parameters of GalNAc-T1 catalyzed transfer to peptides containing acceptor "motifs".

Sequence	*motif N⁰*	K_m *mM*	V_{max} x 10^3 s^{-1}	V_{max}/K_m $M^{-1}s^{-1}$
ELAP-**T**-APPE	1	1.33	6906	5192
ATAA-**T**-AATA	2	1.44	547	380
IGVR-**T**-VAPP	3	23.7	648	27.3
PPKA-**T**-APPP	3	6.97	974	140
APAP-**S**-SPPP	4	1.81	263	145

The cause of the higher *in vitro* specificity toward threonine-containing acceptors is as yet unclear. An examination of kinetic data for the two acceptor pairs PPAS**T**SAPG/PPAS**S**SAPG and PPDAA**T**AAPL/

PPDAASAAPL suggests that if the K_m values reflect binding affinities, then the enhanced catalytic efficiency toward threonine may be due at least partially to a higher binding affinity toward threonine containing acceptors, even though the K_m's for PPAS**T**SAPG and PPAS**S**SAPG are almost identical. The fact that the K_i's for the inhibitor peptides based on the acceptor PPDAA**T**AAPL are very similar to the K_m for PPDAA**S**AAPL, indicates that the increased affinity may be due to some specific interaction of the acceptor subsite with the methyl group of the threonine side chain. (Table 6). Since the substitution of α-amino butyric acid in the position of the acceptor amino acid does not improve the affinity of the peptide for the enzyme (Table 6), it is not the interaction of the subsite with the methyl group of threonine *per se* which is directly responsible for the increased free energy of binding of the peptide to the enzyme. In other words, it is unlikely that the higher efficiency is the result of the enzyme containing a high-affinity binding site specific for a methyl group. Rather, the presence of the methyl group of threonine at the active site must somehow enhance the interaction of the hydroxyl group of threonine with the enzyme. The interaction of the methyl group with a bulky group on the enzyme may create a steric hindrance to the free rotation of the hydroxyl-bearing carbon and thereby promote the formation of a stable hydrogen bond between the hydroxyl group and the enzyme. Whatever the exact cause of threonine specificity, the results from these binding experiments are fully consistent with the hypothesis of cumulative subsite interactions between the enzyme and the acceptor and argue against recognition of specific acceptor motifs. Could it be that the binding of the threonine-containing peptides to the enzyme is stronger because threonine, but not serine, is able to induce in solution a peptide conformation already approximating that of the enzyme-bound peptide? Energy-minimized modeling studies of the acceptor peptide PPDAA**T**AAPL and of homologues where the acceptor threonine is replaced by serine, cysteine, alanine, or 3-hydroxyproline (see Tables 1 and 6), indicate that all these peptides—with or without threonine—have a tendency to form similar structures that contain only a limited number of secondary structural elements. NMR studies of these peptides in solution indicate that even if there may be some weak interactions between amino acids in the amino terminal portion of the peptide PPDAA**T**AAPL, there are no detectable interactions in other efficient acceptor peptides, such as PPAS**T**SAPG (A. Kurosaka, B. Stockman and Å. Elhammer, unpublished observations). Taken together these results are consistent with the proposal that the primary structure is the major determinant of peptide specificity and they indicate that GalNAc-T does not recognize a specific solution secondary structure of the acceptor peptide segment even if the enzyme induces an extended conformation in the peptide [18,81], upon binding.

Table 6. Dissociation constants of peptides acting as competitive inhibitors of GalNAc-T1 catalyzed glycosylations.

Peptide	K_i *mM*
PPDA—**A**—AAPL	1.51
PPDAA—**A**—AAPL	0.98
PPDAA—α**aB**—AAPL	1.58
PPDAA—**C**—AAPL	~4.5
PPDAA—**hyP**—AAPL	21.2

αaB, α-aminobutyric acid
hyP, 3-hydroxyproline

Structural studies on GalNAc-transferase acceptors

The β-carbon in both serine and threonine is able to rotate around the α-carbon-β-carbon bond. This allows the β-hydroxyl group of the enzyme-bound peptide to assume an orientation with respect to the catalytic groups of the enzyme that is conducive to glycosylation. However, free rotation will also allow the hydroxyl group to be orientated in directions that preclude glycosylation. Consequently, the degree of freedom of rotation in the enzyme-bound peptide likely has a significant influence on the catalytic efficiency of an acceptor since selective immobilization of the threonine hydroxyl in an optimal position should increase the rate of glycosyl transfer. We propose that this, in conjunction with the enhanced binding discussed above, is the reason for the differing catalytic efficiencies toward serine and threonine. The kinetic data presented in Tables 1 to 4 are consistent with this hypothesis. Thus, the immobilization of the hydroxyl group by steric hindrance promotes glycosylation, not only in the binding step, but also in the catalytic step. Conversely, the hydroxyl group on serine, which is not subject to this restriction, requires more energy to bind (higher K_d) and even in the bound state it retains some of its mobility which then results in lower catalytic efficiency, hence a lower V_{max}. The fact that the methyl group on threonine does restrict the orientation of the hydroxyl is clearly demonstrated by the poor reactivity of peptides in which the acceptor L-threonine had been replaced by L-*allo*-threonine (Table 1). These peptides are essentially inactive, in spite of the fact that *allo*-threonine contain both the hydroxyl and the methyl groups that make threonine a very efficient acceptor. This result is indeed fully consistent with the proposal that the methyl group is forced by steric hindrance into the same position in *allo*-threonine as in threonine, thereby forcing the hydroxyl of *allo*-threonine to be oriented in a direction away from the catalytic groups of the enzyme. An analogous situation is created by the introduction of 3-hydroxyproline in the acceptor position of the peptide (Table 6). Although this amino acid contains an hydroxyl group on the β-carbon and

although 3-hydroxyproline has the same stereochemistry as threonine, the proline ring structure locks the hydroxyl group in an orientation that is not conducive to GalNAc-T1 catalyzed glycosylation. Consequently, this peptide is merely an inhibitor instead of being an acceptor.

The specificity of GalNAc-transferase appears to be restricted to aliphatic hydroxyl-containing amino acid acceptors. Thus, a peptide containing cysteine in lieu of the acceptor threonine or serine, even when incubated for a long time in the presence of large amounts of recombinant enzyme, failed to produce any measurable amount of glycosylated product, suggesting that GalNAc-T1 is incapable of transferring GalNAc to a thiol group (Table 6). And yet, this peptide does bind to the enzyme with an affinity comparable to that of the corresponding peptides containing serine or alanine at the acceptor site: the K_i for PPDAA**C**AAPL is 4.5 mM as compared to 1 mM for PPDAA**A**AAPL and a K_m of 1.8 mM for PPDAA**S**AAPL. The slightly lower affinity of the cysteine-containing peptide is probably the result of the larger size of the SH group as compared to that of an OH group. Since the cysteine residue in the peptide represents the sulfhydryl equivalent of serine, an amino acid which only has a modest *in vitro* acceptor activity, it is tempting to speculate that an acceptor peptide containing L-β-methylcysteine—the equivalent of L-threonine—may be slowly glycosylated by the enzyme.

Specificity-based design of small molecule inhibitors of GalNAc-transferases; applications in drug discovery research

O-linked oligosaccharide structures have been assigned a range of diverse biological functions [2,82] and they also play key roles in many pathologies. Conditions associated with cell-cell interaction mechanisms, such as host cell-pathogen interactions, leukocyte-endothelial cell interactions, gamete interactions, cancer cell-endothelium cell interaction during metastasis, and cytotoxic effector cell-target cell interactions in immune response reactions, are all prominent illustrations of the important biological roles of O-linked oligosaccharides [2,83]. Some pathological conditions are directly linked to over expression of O-linked oligosaccharide structures. Most human cancer cells, for example, express specific O-linked oligosaccharide antigens at levels that often are directly correlated with malignancy [83]. Two such structures, commonly referred to as the T and Tn antigens, are among the best-characterized and widely recognized cancer markers [84,85]. The importance of O-glycans in carcinogenesis and cancer progression is cogently illustrated by the performance of the cell surface molecule MUC1 in human breast cancer: recent work has shown that over-expression of this molecule, typically at ten times the normal level, promotes tumor growth and also helps human breast cancer cells avoid elimination by the effector cells of the immune system [75,86–91].

With this background, one cannot but conclude that modification or inhibition of O-linked oligosaccharide biosynthesis is a prime target for drug discovery. For example, compounds that inhibit GalNAc-transferase could be used to attenuate the excessive synthesis of cancer-associated mucins and thereby predispose the cancer cells to contact inhibition signals from surrounding tissues as well as to elimination by the immune system. The same type of compounds also could be used to modify saccharide structures expressed in normal cells in such a way that the binding of circulating leukocytes to inflammation sites, or the binding of pathogens to the surface of host cells, could be fully, or at least partially, inhibited. Other possible applications include the treatment of respiratory conditions such as asthma, emphysema, and cystic fibrosis as well as immune/autoimmune diseases. The fact that there are several homologous, but subtly distinct, mammalian GalNAc-transferases [28,31,74]—some of which are expressed in a tissue and/or cell type-specific manner—offers the possibility of developing inhibitors directed at a specific enzyme involved in a particular disease process, thereby avoiding side effects caused by a broad systemic inhibition of O-glycan synthesis.

In preliminary studies evaluating this concept, we were able to identify a number of small molecules capable of inhibiting *in vitro* the GalNAc-T1-catalyzed transfer to synthetic acceptors. Some of these compounds also showed activity in cell cultures. In particular, we found two compounds that are able to inhibit completely the conjugation of O-glycans to the MUC1 precursor proteins in the human breast cancer cell line ZR-75-1. This resulted in a considerable attenuation of the expression of mature MUC1 molecules on the surface of these cells (C.A. Baker, T. Peterson, and Å.P. Elhammer, unpublished observations).

Conclusions and future directions

In summary, the initiation of mucin-type O-linked oligosaccharide biosynthesis is catalyzed by a family of peptide:N-acetylgalactosaminyltransferases with closely overlapping *in vitro* acceptor specificities. These transferases are multisubstrate enzymes with extended active sites containing at least nine subsites that interact cooperatively with a linear segment of at least nine amino acid residues on the acceptor polypeptide. No specific acceptor motifs are recognized by the GalNAc-T's, but efficient acceptor proteins are characterized by a restricted distribution of a limited set of amino acids. Functional acceptor sites must also be accessible on the surface of the protein and extended conformations (β-strand conformation) are preferred. The acceptor specificity of GalNAc-T can be described by an array of specificity parameters characterizing the various amino acids in the nine subsites. These parameters allow

one to use algorithms to predict sites for GalNAc-T catalyzed glycosylation, currently with an accuracy ranging from 70–80%. In spite of significant differences in *in vitro* catalytic efficiencies, the same GalNAc-T enzymes likely catalyze the transfer of N-acetylgalactosamine to both serine and threonine residues *in vivo.* The higher *in vitro* catalytic efficiency observed for threonine as compared to serine is the result of enhanced binding as well as increased reaction velocity. Both of these effects are caused by steric interactions between the active site of the enzyme and the methyl group of threonine. Results from substrate binding studies suggest that GalNAc-T catalyzed transfer of N-acetylgalactosamine to acceptor polypeptides proceeds via an ordered sequential mechanism.

Recent progress in our understanding of the initiation of O-linked oligosaccharide biosynthesis has been impressive: Several transferases catalyzing this reaction have been identified and cloned, the chromosomal localization and genomic organization have been described for three separate enzymes, and a considerable amount of data has been accumulated on the *in vitro* substrate specificities of GalNAc-T. Still, interesting and important questions remain unanswered. With the insight and tools provided by the studies mentioned above it should now be possible to address questions such as optimal reaction conditions, acceptor specificity and enzyme topology *in situ,* as well as the specific function and regulation of the several isozymes identified to date and structural details of the interaction between the enzymes and their substrates.

Acknowledgments

The authors wish to thank Dr. K.C. Chou for calculating the most probable conformations of the acceptor peptides, Ms. C. Bannow for synthesizing the peptide substrates and inhibitors, and Dr. T.J. Raub for critical reading of this manuscript.

References

1 Sadler JE (1984) In *Biosynthesis of Glycoproteins: Formation of O-linked Oligosaccharides* (Ginsburg V, Robbins, PW, eds) pp 199–288. New York: JE Wiley.

2 Schachter H, Brockhausen I (1993) In *Glycoconjugates: Composition, Structure and Function* (Allen HJ, Kisailus EC, eds) pp 263–332. New York: Marcel Dekker, Inc.

3 Roseman S (1970) *Chem Phys Lipids* **5:** 270–97.

4 Seidah NG, Chretien M (1981) *Proc Natl Acad Sci USA* **78:** 4236–40.

5 Titani K, Kumar S, Takio K, Ericsson LH, Wade RD, Ashida K, Walsh KA, Chopek MW, Sadler JE, Fujikawa K (1986) *Biochemistry* **25:** 3171–84.

6 Rohani A, Zannis VI (1988) *J Biol Chem* **263:** 17925–32.

7 Wernette-Hammond ME, Lauer SJ, Corsini A, Walker D, Taylor JM, Rall SC, Jr (1989) *J Biol Chem* **264:** 9094–101.

8 Hayes GR, Enns CA, Lucas JJ (1992) *Glycobiology* **2:** 355–9.

9 Steiner V, Knecht R, Börnsen O, Gassman E, Stone SR, Raschdorf F, Schlaeppi J-M, Maschler R (1992) *Biochemistry* **31:** 2294–98.

10 Shimizu N, Hara H, Sogabe T, Sakai H, Ihara I, Inoue H, Nakamura T, Shimizu S (1992) *Biochim. Biophys. Res. Commun.* **189,** 1329–35.

11 Do S-I, Cummings RD (1992) *Glycobiology* **2,** 345–53.

12 Kim H, Yamaguchi Y, Masuda K, Matsunaga C, Yamamoto K, Irimura T, Takahashi N, Arata Y (1994) *J Biol Chem* **269:** 12345–50.

13 Pirie-Shepherd S, Stevens RD, Andon NL, Enghild JJ, Pizzo SV (1997) *J Biol Chem* **272:** 7408–11.

14 Devine PL, McKenzie FC (1992) *BioEssays* **14:** 619–25.

15 Strous GJ, Dekker J (1992) *Crit Rev Biochem Mol Biol* **27:** 57–92.

16 Van Klinken BJ-W, Dekker J, Büller HA, Einerhand AWC (1995) *Am J Physiol* **269:** G613–27.

17 Jentoft N (1990) *TIBS* **15:** 291–94.

18 Elhammer ÅP, Poorman RA, Brown E, Maggiora LL, Hoogerheide JG, Kézdy FJ (1993) *J Biol Chem* **268:** 10029–38.

19 Hansen JE, Lund O, Engelbrecht J, Bohr H, Nielsen JO, Hansen J-ES, Brunak S (1995) *Biochem J* **308:** 801–13.

20 Butenhof KJ, Gerken TA (1993) *Biochemistry* **32:** 2650–63.

21 Gerken TA, Owens CL, Pasumarthy M (1997) *J Biol Chem* **272:** 9709–19.

22 Neumann GM, Marinaro JA, Bach LA (1998) *Biochemistry* **37:** 6572–85.

23 Marth JD (1996) *Glycobiology* **6,** 701–5.

24 Homa FL, Hollander T, Lehman DJ, Thomsen DR, Elhammer ÅP (1993) *J Biol Chem* **268:** 12609–16.

25 Hagen FK, Van Wuyckhuyse B, Tabak LA (1993) *J Biol Chem* **268:** 18960–65.

26 White T, Bennett EP, Takio K, Sørensen T, Bonding N, Clausen H (1995) *J Biol Chem* **270:** 24156–65.

27 Bennett EP, Hassan H, Clausen H (1996) *J Biol Chem* **271:** 17006–12.

28 Hagen FK, Ten Hagen KG, Beres TM, Balys MM, Van Wuyckhuyse BC, Tabak LA (1997) *J Biol Chem* **272:** 13843–48.

29 Ten Hagen KG, Hagen FK, Balys MM, Beres TM, Van Wuyckhuyse B, Tabak LA (1998) *J Biol Chem* **273:** 27749–54.

30 Wandall HH, Hassan H, Mirgorodskaya E, Kristensen AK, Roepstorff P, Bennett EP, Nielsen PA, Hollingsworth MA, Burchell J, Taylor-Papadimitriou J, Clausen H (1997) *J Biol Chem* **272:** 23503–14.

31 Clausen H, Bennett EP (1996) *Glycobiology* **6:** 635–46.

32 Tooze SA, Tooze J, Warren G (1988) *J Cell Biol* **106:** 1475–87.

33 Pathak RK, Merkle RK, Cummings RD, Goldstein JL, Brown MS, Anderson RGW (1988) *J Cell Biol* **106:** 1831–41.

34 Schweizer A, Clausen H, van Meer G, Hauri H-P (1994) *J Biol Chem* **269:** 4035–41.

35 Johnson DC, Spear PG (1983) *Cell* **32:** 987–97.

36 Roth J (1984) *J Cell Biol* **98,** 399–406.

37 Elhammer Å, Kornfeld S (1984) *J Cell Biol* **98,** 327–331.

38 Deschuyteneer M, Eckhardt AE, Roth J, Hill RL (1988) *J Biol Chem* **263:** 2452–59.

39 Ulmer JB, Palade GE (1989) *Proc Natl Acad Sci* **86:** 6992–96.

40 Piller V, Piller F, Fukuda M (1990) *J Biol Chem* **265:** 9264–71.

41 Roth J, Wang Y, Eckhardt AE, Hill RL (1994) *Proc Natl Acad Sci* **91:** 8935–39.

42 Abeijon C, Hirschberg CB (1987) *J Biol Chem* **262:** 4153–59.
43 Abeijon C, Mandon EC, Hirschberg CB (1997) *TIBS* **22:** 203–7.
44 Nehrke K, Ten Hagen KG, Hagen FK, Tabak LA (1997) *Glycobiology* **7:** 1053–60.
45 Hagen FK, Van Wuyckhuyse B, Tabak LA (1993) *J Biol Chem* **268:** 18960–65.
46 Kearns AE, Campbell SC, Westley J, Schwartz NB (1991) *Biochemistry* **30:** 7477–83.
47 Hagopian A, Eylar EH (1968) *Arch Biochem Biophys* **128:** 422–33.
48 Hagopian A, Westall FC, Whitehead JS, Eylar EH (1971) *J Biol Chem* **246:** 2519–23.
49 Hill HD, Schwyzer M, Steinman HM, Hill RL (1977) *J Biol Chem* **252:** 3799–804.
50 Young JD, Tsuchiya D, Sandlin DE, Holroyde MJ (1979) *Biochemistry* **18:** 4444–48.
51 Briand JP, Andrews SP Jr., Cahill E, Conway NA, Young JD (1981) *J Biol Chem* **256:** 12205–207.
52 Cruz TF, Moscarello MA (1983) *Biochim Biophys Acta* **760:** 403–10.
53 Wang Y, Abernethy JL, Eckhardt AE, Hill RL (1992) *J Biol Chem* **267:** 12709–16.
54 O'Connell B, Tabak LA, Ramasubbu N (1991) *Biochem Biophys Res Commun* **180:** 1024–30.
55 O'Connel BC, Hagen FK, Tabak LA (1992) *J Biol Chem* **267:** 25010–18.
56 O'Connel BC, Tabak LA (1993) *J Dent Res* **72:** 1554–58.
57 Wang Y, Agrwal N, Eckhardt AE, Stevens RD, Hill RL (1993) *J Biol Chem* **268:** 22979–83.
58 Nishimori I, Johnson NR, Sanderso SD, Perini F, Mountjoy K, Cerny RL, Gross ML, Hollingsworth MA (1994) *J Biol Chem* **269:** 16123–30.
59 Nishimori I, Perini F, Mountjoy KP, Sanderson SD, Johnson N, Cerny RL, Gross ML, Fontenot JD, Hollingsworth MA (1994) *Cancer Res* **54:** 3738–44.
60 Wragg S, Hagen FK, Tabak LA (1995) *J Biol Chem* **270:** 16947–54.
61 Stadie TRE, Chai W, Lawson AM, Byfield PGH, Hanisch F-G (1995) *Eur J Biochem* **229:** 140–47.
62 Brockhausen I, Toki D, Brockhausen J, Peters S, Bielfeldt T, Kleen A, Paulsen H, Meldal M, Hagen F, Tabak LA (1996) *Glycoconj J* **13:** 849–56.
63 Yoshida A, Suzuki M, Ikenaga H, Takeuchi M (1997) *J Biol Chem* **272:** 16884–88.
64 Sugahara T, Pixley MR, Fares F, Boime I (1996) *J Biol Chem* **271:** 20797–804.
65 Kornfeld R, Kornfeld S (1985) *Ann Rev Biochem* **54:** 631–64.
66 Gooley AA, Classon BJ, Marschalek R, Williams KL (1991) *Biochem Biophys Res Commun* **178:** 1194–201
67 Pisano A, Redmond JW, Williams KL, Gooley AA (1993) *Glycobiology* **3:** 429–35.
68 Pisano A, Packer NH, Redmond JW, Williams KL, Gooley AA (1994) *Glycobiology* **4:** 837–44.
69 Gooley AA, Williams KL (1994) *Glycobiology* **4:** 413–17.
70 Sugahara T, Pixley MR, Fares F, Boime I (1996) *J Biol Chem* **271:** 20797–804.
71 Chou K-C, Zhang C-T, Kézdy FJ, Poorman RA (1995) *Proteins* **21:** 118–26.
72 Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S (1997) *Glycoconj J* **15:** 115–30.
73 Nehrke K, Hagen FK, Tabak LA (1996) *J Biol Chem* **271:** 7061–65.
74 Hagen FK, Nehrke K (1998) *J Biol Chem* **273:** 8268–77.
75 Hilkens J, Buijs F (1988) *J Biol Chem* **263:** 4215–22.
76 Dekker J, Van Beurden-Lamers WMO, Strous GJ (1989) *J Biol Chem* **264:** 10431–37.
77 Hull SR, Sugarman ED, Spielman J, Carraway KL (1991) *J Biol Chem* **266:** 13580–86.
78 McCool DJ, Forstner JF, Forstner GG (1994) *Biochem J* **302:** 111–18.
79 Delorme, Lorenzini T, Griffin J, Martin F, Jacobsen F, Boone T, Elliott S (1992) *Biochemistry* **31:** 9871–76.
80 Elliott S, Bartley T, Delorme E, Derby P, Hunt R, Lorenzini T, Parker V, Rohde MF, Stoney K (1994) *Biochemistry* **33:** 11237–45.
81 Callaghan Rose M, Voter WA, Sage H, Brown CF, Kaufman B (1984) *J Biol Chem* **259:** 3167–72.
82 Hounsell EF, Davies MJ, Renouf DV (1996) *Glycoconj J* **13:** 19–26.
83 Brockhausen I (1993) *Crit Rev Clin Lab Sci* **30:** 65–151.
84 Springer GF (1984) *Science* **224:** 1198–206.
85 Takahashi HK, Metoki R, Hakamori S (1988) *Cancer Res* **48:** 4361–67.
86 Gendler, SJ, Lancaster CA, Taylor-Papadimitriou J, Duhig T, Peat N, Burchell J, Pemberton L, Lalani E-N, Wilson D (1990) *J Biol Chem* **265:** 15286–93.
87 Devine PL, McKenzie FC (1992) *BioEssays* **14:** 619–25.
88 van de Wiel-van de Kemenade E, Ligtenberg MJL, de Boer AJ, Buijs F, Vos HL, Melief CJM, Hilkens J, Figdor CG (1993) *J Immunol* **151:** 767–76.
89 Verma M, Davidson EA (1994) *Glycoconj J* **11:** 172–79.
90 Hilkens J, Wesseling J, Vos HL, Litvinov SL, Boer M, van der Valk S, Calafat J, Patriarca C, van der Wiel-van der Kemenade E, Figdor C (1995) In *Biochemistry of Cell Membranes* (Papa S, Tager JM, eds) pp 259–71. Basel/Switzerland: Birkhäuser Verlag.
91 Gimmi CD, Morrison BW, Mainpric BA, Gribben JG, Boussiotis VA, Freeman GJ, Park SYL, Wanatabe M, Gong JL, Hayes DF, Kufe DW, Nadler LM (1966) *Nature Med* **2:** 1367–70.

Accepted 12 February 1999.