Mucin Core O-Glycosylation Is Modulated by Neighboring Residue Glycosylation Status

KINETIC MODELING OF THE SITE-SPECIFIC GLYCOSYLATION OF THE APO-PORCINE SUBMAXILLARY MUCIN TANDEM REPEAT BY UDP-GalNAc:POLYPEPTIDE N-ACETYLGLACTOSAMINYLTRANSFERASES T1 AND T2

Thomas A. Gerken‡§, Jiexin Zhang‡, Jessica Levine, and Ake Elhammer¶

From the Departments of Pediatrics and Biochemistry, W. A. Bernbaum Center for Cystic Fibrosis Research and University Hospitals Research Institute, Case Western Reserve University School of Medicine, Cleveland, Ohio 44106 and ¶Pharmacia Corporation, Kalamazoo, Michigan 49001

The influence of peptide sequence and environment on the initiation and elongation of mucin O-glycosylation is not well understood. The in vivo glycosylation pattern of the porcine submaxillary gland mucin (PSM) tandem repeat containing 31 O-glycosylation sites (Gerken, T. A., Gilmore, M., and Zhang, J. (2002) J. Biol. Chem. 277, 7758–7761) reveals a weak inverse correlation with hydroxyamino acid density (and by inference the density of glycosylation) with the extent of GalNAc glycosylation and core-1 substitution. We now report the time course of the in vitro glycosylation of the apoPSM tandem repeat by recombinant UDP-GalNAc:polypeptide α-GalNAc transferases (ppGalNAc transferase) T1 and T2 that confirm these findings. A wide range of glycosylation rates are found, with several residues showing apparent plateaus in glycosylation. An adjustable kinetic model that reduces the first-order rate constants proportional to neighboring glycosylation status, plus or minus three residues of the site of glycosylation, was found to reasonably reproduce the experimental rate data for both transferases, including apparent plateaus in glycosylation. The unique, transferase-specific, positional weighting constants reveal information on the peptide/glycopeptide recognition site for each transferase. Both transferases displayed high sensitivities to neighboring Ser/Thr glycosylation, whereas ppGalNAc T2 displayed additional high sensitivities to the presence of nonglycosylated Ser/Thr residues. This is the first demonstration of the ability to model mucin O-glycosylation kinetics, confirming that under the appropriate conditions neighboring glycosylation status can be a significant factor modulating the first step of mucin O-glycan biosynthesis.

A wide range of secreted and membrane-associated proteins are O-glycosylated at serine and threonine by glycans linked through N-acetylgalactosamine (GalNAc). A significant number of these glycoproteins contain heavily O-glycosylated domains rich in Ser and Thr of which mucus glycoproteins, commonly called mucins, represent a major class (1). These heavily O-glycosylated domains typically contain 20–30% Ser and Thr, are up to 80% carbohydrate by weight, and are commonly made up of tandemly repeated peptide sequences. Many important biological processes including the protection of epithelial cell surfaces, the immune response, adhesion, inflammation, and tumor genesis (2–9) appear to be afforded and/or modulated by mucins or glycoproteins containing such mucin-like domains. Protein O-glycosylation may also play important developmental roles (10, 11).

In the Golgi the transfer of GalNAc to Ser/Thr residues by UDP-GalNAc:polypeptide α-GalNAc transferase (ppGalNAc transferase) represents the first step in O-glycan synthesis. Twelve members of the mammalian ppGalNAc transferase family have been described to date, ppGalNAc T1–T12 (9–19). Homologous ppGalNAc transferases have been described in Drosophila and Caenorhabditis elegans (10, 20). Although not well characterized, it is accepted that peptide substrate specificities can vary among the different family members (10, 13, 21–25). In addition, the most characterized transferases (principally ppGalNAc T1 through T4) have activities and specificities that seem to be unpredictably altered by prior peptide substrate glycosylation (26–30), whereas ppGalNAc transferases T7 and T102 show an apparent absolute requirement for prior GalNAc addition for activity (9, 16, 31). The expression of the various ppGalNAc transferase family members having different peptide and/or glycopeptide specificities therefore represents the initial step in the regulation or modulation of O-glycan biosynthesis. Subsequent elongation of O-linked glycans proceeds by the stepwise addition of single sugar residues via a series of substrate-specific Golgi resident transferases (32). Depending on the initial and subsequent substitutions on the GalNAc residue (which may be further modulated by peptide sequence and glycosylation state), a bewildering array of O-linked structures is possible (32).

Central to our understanding of the role and regulation of mucin-type O-glycosylation is the quantitative determination through N-acetylgalactosamine (GalNAc).1 A significant number of these glycoproteins contain heavily O-glycosylated domains rich in Ser and Thr of which mucus glycoproteins, commonly called mucins, represent a major class (1). These heavily O-glycosylated domains typically contain 20–30% Ser and Thr, are up to 80% carbohydrate by weight, and are commonly made up of tandemly repeated peptide sequences. Many important biological processes including the protection of epithelial cell surfaces, the immune response, adhesion, inflammation, and tumor genesis (2–9) appear to be afforded and/or modulated by mucins or glycoproteins containing such mucin-like domains. Protein O-glycosylation may also play important developmental roles (10, 11).

In the Golgi the transfer of GalNAc to Ser/Thr residues by UDP-GalNAc:polypeptide α-GalNAc transferase (ppGalNAc transferase) represents the first step in O-glycan synthesis. Twelve members of the mammalian ppGalNAc transferase family have been described to date, ppGalNAc T1–T12 (9–19). Homologous ppGalNAc transferases have been described in Drosophila and Caenorhabditis elegans (10, 20). Although not well characterized, it is accepted that peptide substrate specificities can vary among the different family members (10, 13, 21–25). In addition, the most characterized transferases (principally ppGalNAc T1 through T4) have activities and specificities that seem to be unpredictably altered by prior peptide substrate glycosylation (26–30), whereas ppGalNAc transferases T7 and T102 show an apparent absolute requirement for prior GalNAc addition for activity (9, 16, 31). The expression of the various ppGalNAc transferase family members having different peptide and/or glycopeptide specificities therefore represents the initial step in the regulation or modulation of O-glycan biosynthesis. Subsequent elongation of O-linked glycans proceeds by the stepwise addition of single sugar residues via a series of substrate-specific Golgi resident transferases (32). Depending on the initial and subsequent substitutions on the GalNAc residue (which may be further modulated by peptide sequence and glycosylation state), a bewildering array of O-linked structures is possible (32).

Central to our understanding of the role and regulation of mucin-type O-glycosylation is the quantitative determination

1 The abbreviations used are: GalNAc, N-acetylgalactosamine; pp-GalNAc T, UDP-GalNAc:polypeptide α-GalNAc transferase; PSM, porcine submaxillary mucin; r2, least square correlation coefficient.

2 Note that there are two different mammalian ppGalNAc T9s described in the literature: Toha and co-workers (18) and Ten Hagen and co-workers (9). The latter should perhaps be referred to as ppGalNAc T10.
of the site-specific glycosylation pattern of mucins and mucin-like domains and attempting to relate these patterns to the nature of the surrounding polypeptide sequence. A number of statistically based approaches have been reported that have attempted to predict sites of O-glycosylation; however, these approaches are at best semiquantitative and are thus not capable of predicting the extent of site-specific glycosylation (33–35). Presently no predictive approaches exist for estimating site-specific oligosaccharide side chain structures.

Many studies have been performed both in vivo and in vitro examining the glycosylation of specific peptide sequences with the goals of characterizing the peptide specificities of the various GalNAC transferases (26, 36–40). To date only the ppGalNAC T1 has been extensively and systematically characterized with respect to peptide sequence (36). The results of these and other in vitro studies and the above statistical analyses all point to the roles of neighboring proline, serine, and threonine residues for enhancing the probability of O-glycosylation. The effects of peptide O-glycosylation on ppGalNAC transferase activity have unfortunately failed to provide any generalized rules for O-glycosylation (26–30).

With the goal of understanding the influence of peptide sequence and structure on mucin O-glycosylation, our laboratory has been systematically characterizing the in vivo site-specific glycosylation pattern of each of the 31 glycosylated residues in the 81-residue tandem repeat of the porcine submaxillary mucin (PSM) (41, 42) (see Fig. 2, panel C, for the tryptic tandem repeat sequence of PSM (43, 44)). Recently we have reported (45) the mono-(α-GalNAc-O-Ser/Thr), di-(β-Gal-1,3-α-GalNAc-O-Ser/Thr), and tri-(α-Fuc-1,2-β-Gal-1,3-α-GalNAc-O-Ser/Thr) saccharide distributions at nearly each individual glycosylation site of the PSM tandem repeats isolated from a group of A blood group minus animals. Our analysis of the glycosylation pattern of these mucins was found to support our earlier suggestions that the O-linked glycan side chain structures and side chain lengths of mucin appear to be modulated in vivo by the density of neighboring, presumably partially glycosylated, hydroxyaminic acid residues in the polypeptide sequence. This effect was detected initially for the core-1 structures, formed by the transfer of β-Gal to GalNAc by the core-1 β-galactosyltransferase (41) and more recently for the peptide GalNAc residues (45).

In this report we have extended our studies to the characterization of the in vitro glycosylation of the apoPSM tandem repeat domain by purified recombinant ppGalNAC transferases T1 and T2. The results show a wide range of rates of GalNAc incorporation with Thr residues incorporating GalNAc at a significantly higher rate than Ser residues. Kinetic models taking into account the reduction of the Ser and Thr first-order rate constants proportionally to the glycosylation status of neighboring residues were found to reasonably reproduce the experimental site-specific rates of GalNAc incorporation for both ppGalNAC transferases. These findings unambiguously confirm the important role that neighboring group glycosylation plays in modulating the initial step of mucin O-glycosylation. Furthermore, the positional weighting constants derived from the models may be used to infer information on the peptide/glycopeptide recognition sites on the different ppGalNAC transferases. This is the first report demonstrating the ability to kinetically model the site-specific glycosylation pattern of a heavily O-glycosylated mucin-like domain.

MATERIALS AND METHODS

Mucin Substrate Preparation—Oligomeric PSM tandem repeat glycosylated domains were obtained after trypsinization and gel filtration chromatography of the reduced and carboxymethylated mucin as described previously (41, 42). Domains were fully deglycosylated by mild trifluoroacetate anilide treatment followed by periodate oxidation and alkaline elimination as described (46). Carbon-13 NMR spectroscopy was used to confirm the complete removal of carbohydrate. Apo-tandem repeat domains were fractionated on Sephadryl S200, 50 mM NaH2CO3, pH 8.5, buffer, and the relatively broad apomucin peak split into 4 fractions. The highest molecular weight fraction, representing ~1/4 of the preparation, was utilized as an acceptor peptide. In some studies apomucin was further freed of potential contaminating proteases by performic acid oxidation and alkaline elimination (46). Carbon-13 NMR spectroscopy was used to confirm the complete oxidation and alkaline elimination of the full-length 81-residue tandem repeat glycopeptide of PSM as obtained by the above biotinylation procedure. For all time points, except for the 80-h time point with ppGalNAC T2, a unique set of N- and C-terminal glycopeptides (better than 80% a single sequence) could be obtained for sequence analysis. For the 80-h ppGalNAC T2 experiment, the full-length 81-residue tryptic tandem repeat was lost to extensive Arg-N terminal protease (45)–epoxysuccinyl-L-leucylamido(4-hydroxyphenyl)alanine (Bestatin) and leupeptin, and aprotinin, were used. Reaction mixtures were incubated at 37 °C and were subsequently transferred to 3.5-kDa molecular weight cut-off dialysis membranes (Spectrum Laboratories, Rancho Dominguez, CA, or Pierce) for dialysis at 4 °C against 2–3 changes of 75 mM acetate buffer (pH 4.5 with NH4OH) to remove free UDP inhibitor. GalNAc transferase was reinitiated at 37 °C by adding again both protease inhibitors and UDP-GalNAc (to 5 mM). This procedure was repeated until the desired net reaction time was reached (up to ~250 h) while adding fresh transferase (11 mg/ml) at approximately the half-lives of each transferase, determined under similar reaction conditions (see the Supplemental Material). After the final incubation the sample was dialyzed against water, lyophilized, and fractionated on Sephacryl S200 in 50 mM ammonium bicarbonate, pH 8.5, buffer, and the relatively broad apo-tandem repeat domains were fractionated on Sephacryl S200 in 50 mM ammonium bicarbonate, pH 8.5, or in later experiments in 50 mM acetate buffer (pH 4.5 with NH4OH) to reduce microbially derived proteolytic cleavage of the tandem repeat domain. Intact and partially cleaved PSM tandem repeat domain glycopeptides were pooled separately.

PSM Tandem Repeat Glycopeptide Isolation—The partially glycosylated PSM 81-residue tandem repeat glycopeptide was obtained from S200 chromatography after trypsinolysis as described previously (41, 42). The C-terminal portion of the 81-residue tandem repeat glycopeptide (residues 39–78) was obtained after N-terminal biotinylation of the 81-residue tandem repeat, digestion with protease Glu-C, and passage through an immobilized avidin column (Pierce) as described previously (43). Commonly after lengthy incubations with transferase, the oligosaccharide tandem repeat domain exhibited various degrees of cleavage by contaminating proteases. Such cleavage was significantly reduced with the use of the commercial protease mixtures. These glycopeptides, which were the size of ~80 and 40 residues on gel filtration, were characterized by N-terminal Edman amino acid sequencing. The most common cleavage was found to occur at the N terminus of Arg, which gave predominantly the ~81-residue tandem repeat glycopeptide starting at Arg78 and two overlapping ~40-residue glycopeptides beginning at Arg24 and Arg34. Attempts to separate the latter two overlapping 40-residue glycopeptides by reverse phase high pressure liquid chromatography were unsuccessful. The C-terminal peptide, residues 39–78, of the full-length Arg24 tandem repeat glycopeptide was obtained by the above biotinylation procedure. For all time points, except for the 80-h time point with ppGalNAC T2, a unique set of N- and C-terminal glycopeptides (better than 80% a single sequence) could be obtained for sequence analysis. For the 80-h ppGalNAC T2 experiment, the full-length 81-residue tryptic tandem repeat was lost to extensive Arg-N terminus of Arg, which gave predominantly the ~81-residue tandem repeat glycopeptide starting at Arg78 and two overlapping ~40-residue glycopeptides beginning at Arg24 and Arg34. Attempts to separate the latter two overlapping 40-residue glycopeptides by reverse phase high pressure liquid chromatography were unsuccessful. The C-terminal peptide, residues 39–78, of the full-length Arg24 tandem repeat glycopeptide was obtained by the above biotinylation procedure. For all time points, except for the 80-h time point with ppGalNAC T2, a unique set of N- and C-terminal glycopeptides (better than 80% a single sequence) could be obtained for sequence analysis. For the 80-h ppGalNAC T2 experiment, the full-length 81-residue tryptic tandem repeat was lost to extensive Arg-N terminus of Arg, which gave predominantly the ~81-residue tandem repeat glycopeptide starting at Arg78 and two overlapping ~40-residue glycopeptides beginning at Arg24 and Arg34. Attempts to separate the latter two overlapping 40-residue glycopeptides by reverse phase high pressure liquid chromatography were unsuccessful. The C-terminal peptide, residues 39–78, of the full-length Arg24 tandem repeat glycopeptide was obtained by the above biotinylation procedure. For all time points, except for the 80-h time point with ppGalNAC T2, a unique set of N- and C-terminal glycopeptides (better than 80% a single sequence) could be obtained for sequence analysis. For the 80-h ppGalNAC T2 experiment, the full-length 81-residue tryptic tandem repeat was lost to extensive Arg-N
proteolysis prior to obtaining its C-terminal sequence by the biotinyl-
ation procedure. The glycosylation of its C-terminal fragment was de-
ered mathematically using the multiple sequence data from the com-
bined 40-residue peptides (beginning with Ile1 and Arg14) and from the
original full-length tryptic tandem repeat. In other preparations, cleav-
age sites at the N terminus of Val and the C terminus of Glu49 were
observed, the latter because of protease Glu-C contamination. This
latter cleavage allowed the determination of the glycosylation of Thr79
and Ser80 which are typically very difficult to obtain (see Fig. S3 in the
Supplemental Material). In all experiments the observed site-specific
glycosylation of the partially proteolyzed tandem repeat sequences was
found to be essentially the same as the glycosylation for the full-length
tandem repeat. Subsequent experiments revealed the Sephacryl S200
column in (NH4)2CO3 buffer as the source of the Arg-N proteolytic
activity presumably due to microbial contamination. Changing the col-
umn buffer to 50 mM acetic acid, pH 4.5, has eliminated the Arg-N
proteolytic activity (data not shown).

Amino Acid Sequencing—Pulsed liquid phase Edman degradation
amino acid sequencing was performed on an Applied Biosystems Pro-
cise 494 protein sequencer (Applied Biosystems, Foster City, CA), and
site-specific glycosylation was determined as described previously (41,
42, 45). The reproducibility between sequencing runs is relatively good,
with standard deviations typically in the range of 5% values. Typically
two or more sequence determinations were performed for each isolated
protein and proteolytic activity (data not shown).

Numerical Simulation of Site-specific Glycosylation Kinetics—Mod-
eling of the site-specific glycosylation kinetics taking into account
neighboring group glycosylation was performed using the Lotus 123
spread sheet software release 9.5 (Lotus Development Co., Cambridge,
MA). In the model the instantaneous relative rate of glycosylation of
residue i, \( d[OG]/dt \), is defined as shown in Equation 1,

\[
d[OG]/dt = \kappa_{ser, Thr} f(OG + OH, OH)
\]

where [OH], is the free, nonglycosylated fraction of hydroxyamino acid
residues, and \( \kappa_{ser, Thr} \) is the first-order rate constant for the glycosylation
of residue i, specific to whether it is a Ser or Thr residue. Values for \( \kappa_{ser} \)
and \( \kappa_{Thr} \) are global values and are fixed throughout the simulation.

The function defined as \( f(OG + OH) \), represents a unitless multiplier that
decreases the rate constant commensurate to the current local glycos-
ylation status, referred to as (OG + OH), of the Ser and Thr residues
neighboring residue i. Values for (OG + OH) are obtained from the
weighted summation of the glycosylation status of neighboring residues
as described below. Out of necessity the values for \( f(OG + OH) \), range from 1.0 to
0, representing no alteration of the rate constant to the full suppres-
sion of the rate constant, respectively, based on the modified
exponential function described below. In this manner, the rate constant
may be continuously modulated as a function of local glycosylation
status throughout the course of the simulation.

Values for the local glycosylation status, \( f(OG + OH) \), are defined as
the weighted summation of the neighboring hydroxyamino acid resi-
due's fraction of glycosylation \( [OG]_i \), and fraction free of glycosylation
\( ([OH]_i) \), summed over \( n = 3 \) to \( 3 \) residues adjacent to residue i, as
shown in Equations 2 and 3,

\[
OG = W_{OG} [OG]_i + W_{OG-1} [OG]_{i-1} + W_{OG-2} [OG]_{i-2} + W_{OG-3} [OG]_{i-3}
+ W_{OG+1} [OG]_{i+1} + W_{OG+2} [OG]_{i+2} + W_{OG+3} [OG]_{i+3}
\]

and

\[
OH = W_{OH} [OH]_i + W_{OH-1} [OH]_{i-1} + W_{OH-2} [OH]_{i-2} + W_{OH-3} [OH]_{i-3}
+ W_{OH+1} [OH]_{i+1} + W_{OH+2} [OH]_{i+2} + W_{OH+3} [OH]_{i+3}
\]

In the above equations, the \( W_{OG} \), and \( W_{OH} \), values represent global
positioning weighting coefficients that can be adjusted to reflect the unique
sensitivities of each transferase to local sequence. Values of \( W_{OG} \),
range from 0 to 1. \( W_{OG} \), and \( W_{OH} \), coefficients for non-hydroxyamino acid
residues are defined as zero. Fractional \( [OG]_i \), and \( [OH]_i \), values also range
between 0 and 1 and are determined numerically from the previous time
step of the simulation as described below. The maximum attainable value
for each OG or OH glycosylation status summation is 6. (Note that the
glycosylation status of residue i does not contribute to the determination
of \( [OG + OH] \) value.) This formalism has the flexibility to account for
the inhibitory effects of both glycosylated and non-glycosylated neighbor-
ing hydroxyamino acid residues based on the individual values of the
global positioning weighting coefficients \( W_{OG} \), and \( W_{OH} \).

The \( f(OG + OH) \), function converts the glycosylation status value
\( [OG + OH] \) into a rate constant multiplier with limits of 1 and 0. The
function must have the characteristics so that in the event of the full
 glycosylation of any single neighboring residue the near full inhibition
of glycosylation may be permitted, i.e. as \( [OG + OH] \) approaches 1, \( f(OG + OH) \) approaches 0. To meet this criteria a modified exponential
function, Equation 4 below and plotted in Fig. 1, was devised to satisfy
the above limits. Because the maximum value of the local glycosylation
status, \( [OG + OH] \), will vary depending on the number of Ser and Thr
residues in the local sequence, a normalized inverse linear function will
not suffice because it will not equally weight the glycosylation of individ-
ual Ser or Thr residues in sequences with different numbers of Ser and
Thr residues. The \( f(OG + OH) \), function is defined as follows:

\[
f(OG + OH) = e^{-([OG + OH] C1_{ser} + [OG + OH] C1_{Thr}) (1 + [OG + OH] C2)}
\]

where \( C1_{ser} \), and \( C2 \) are global parameters whose values can be
adjusted to optimize the fit of the simulation to the experimental data.
The term containing \( C1_{ser} \) is included to further linearize the curve (note
that when \( C2 = 0 \) a pure exponential function results). As shown by the
black curve in Fig. 1, values for \( C1 \) of 0.35 and \( C2 \) of 2.2 produce a
ideal shaped curve with an initial slope of \( -1.0 \) and an approximately
95% reduction in value at \( [OG + OH] \) values of \( 0.1 \). The function as defined
by these constants was found to work well for modeling the Thr residue
glycosylation data for both transferases (shown below). It was found after
some experimentation, however, that significantly improved results
were obtained in the modeling of Ser residues when the \( C1_{ser} \) value was
reduced to 0.25, giving the gray curve shown in Fig. 1. This curve is
somewhat steeper than that for Thr having an initial slope of approxi-
mately \(-1.5 \) and giving an \(-95\% \) reduction in value at \( [OG + OH] \)
values of \(-1 \). As for the Thr residues, this function as parameterized
was found to satisfactorily model the Ser data for both ppGalNAc T1
and T2.

The glycosylation kinetics of each transferase was simulated numeri-
cally by using Equation 5 utilizing 360 time steps, \( \Delta t \). For ppGalNAc T1
each step represented 0.25 h, while for ppGalNAc T2 each step repre-
sented 0.9 h. This gave effective simulation times of 90 and 324 h,
respectively. In the equation \( [OG]_i \), and \( [OH]_i \), represent the frac-
tional glycosylation of residue i at time \( t \) and at time \( t + \Delta t \), respectively.

\[
[OG]_{i,t+\Delta t} = \Delta t [OG]_{i,t} f(OG + OH) [OH]_i, + [OG]_i \]

In this model there are a total of 17 adjustable variables. These consist of 12 positional weighting coefficients \( W_{OG} \), and \( W_{OH} \), two Ser
and Thr-specific rate constants \( k_{ser} \) and \( k_{Thr} \) (with units of fractional
residue glycosylation h\(^{-1}\)), and the \( f(OG + OH) \) constants \( C1_{ser} \), \( C1_{Thr} \),
(\( C2 \)).
and C2.3. The implementation and optimization of the model involved the manual adjustment of the above variables until a satisfactory fit (visually, by least square correlation coefficient, $r^2$, and by standard deviation determination) was obtained for the glycosylation of the majority of Ser and Thr residues using all experimental time points.

RESULTS

Optimization of apoPSM Glycosylation by ppGalNAc T1 and T2—Native mucin is highly O-glycosylated by GalNAc (~70%); therefore, it was deemed necessary in these studies to follow the in vitro glycosylation to similar high levels of glycosylation, rather than simply obtaining the initial rates of glycosylation as is typically performed. Only from such high glycosylation studies was it anticipated that the full effects of neighboring glycosylation would be revealed on the glycosylation kinetics. Preliminary studies described in the Supplemental Material indicate that GalNAc incorporation into apoPSM by ppGalNAc T1 reaches a plateau after 5–8 h of incubation (see Fig. S1, panels A and B) and that this is due to both UDP-GalNAc depletion and UDP product inhibition resulting from the competing UDP-GalNAc hydrolyase activity of the transferase. Therefore, to reach high levels of GalNAc incorporation, repetitive incubations were performed, followed by overnight dialysis as described under “Materials and Methods.” Incorporation was reinitiated by the addition of fresh UDP-GalNAc and protease inhibitors, and the cycle was repeated. Because the half-life of ppGalNAc T1 is ~5 days at 37 °C (see Fig. S1, panel C, in the Supplemental Material), every 5th day one-half of the initial ppGalNAc T1 was re-added to the incubation mix. Net incubation times of ~5, 15, 35, and 70 h at 37 °C were obtained for ppGalNAc T1. Carbon-13 NMR spectra of the mucin products prior to trypsinolysis clearly indicate the incorporation of GalNAc and demonstrate that Thr residues are glycosylated at a greater rate than Ser residues (see Fig. S2 in the Supplemental Material). ppGalNAc T2 was shown to maintain linear GalNAc incorporation over at least a 24-h incubation period at 37 °C showing no apparent UDP inhibition (Fig. S1, panels A and B); therefore, repetitive incubations of 24–40 h were performed as described for ppGalNAc T1. Fresh ppGalNAc T2 was added every second dialysis to compensate for the ~2-day half-life of ppGalNAc T2 (Fig. S1, panel C). Because of its roughly one-fifth lower activity relative to ppGalNAc T1, net incubation times of ~40, 80, and 250 h at 37 °C were obtained. Carbon-13 NMR spectra confirmed the incorporation of GalNAc (data not shown).

Site-specific Glycosylation—The site-specific glycosylation patterns obtained at each time point for both ppGalNAc T1 and T2 are plotted with respect to sequence position in Fig. 2 and by hydroxyamino acid residue type as shown in Fig. S4 of the Supplemental Material (also see Table I of the Supplemental Material for a tabulation of the data). Also plotted in each figure are previously determined in vivo GalNAc glycosylation for each residue (the right-most light gray bar in each cluster) (45). A wide range of apparent glycosylation rates are displayed by both transferases, with ppGalNAc T1 glycosylating many more residues than ppGalNAc T2. The longest incubation with ppGalNAc T1 (70 h) resulted in an average residue glycosylation of 49%, whereas the longest incubation with ppGalNAc T2 (250 h) gave an average residue glycosylation of 20%. The figures reveal that Thr residues are typically more highly glycosylated than Ser residues, 78 vs 36% and 35 vs 12%

Note that in the implementation of the model, the possibility for separate positional weighting values for Ser and Thr residues and the inclusion of residues plus and minus 4 from the site of glycosylation were allowed. It has been found, however, that these additional variables do not typically improve the simulation and are therefore not included in the present work.

for ppGalNAc T1 (70 h of incubation) and T2 (250 h of incubation), respectively, although a number of Ser residues can attain high degrees of glycosylation for both transferases. To examine the extent that the glycosylation pattern of each transferase may correspond with the native in vivo glycosylation pattern, the plots in Fig. 3, panels A–D, were obtained. For ppGalNAc T1 there are good correlations between the averaged residue-specific glycosylation at the 70-h time point with the native in vivo glycosylation for both Ser and Thr residues, as shown in Fig. 3, panels A and B ($r^2 = 0.45$, $p = 0.002$ for Ser and $r^2 = 0.65$, $p = 0.002$ for Thr). Very weak, if any, correlation is observed for the 250-h glycosylation by ppGalNAc T2 with the native glycosylation as shown by Fig. 3, panels C and D (see Fig. 3 legend for statistics). These observations and the plots in Fig. 2 and Fig. S4 in the Supplemental Material suggest that ppGalNAc T1 may be a major contributor to the in vivo glycosylation of the porcine submaxillary gland mucin, whereas ppGalNAc T2 may play a minor role. This is consistent with immunohistological studies of the porcine salivary gland, demonstrating that ppGalNAc T1 is expressed at a significantly higher level than ppGalNAc T2 using monoclonal antibodies against the human transferases (51). Because the porcine, bovine, and human ppGalNAc T1 transferases have higher than 98% sequence homology, they are expected to have virtually identical enzymatic and immunological properties. It is likely that those sites not highly glycosylated by ppGalNAc T1 in vitro but highly glycosylated in vivo may be glycosylated in vivo by other ppGalNAc transferases with different peptide/glycopeptide specificities (13).

Previous studies from our laboratory (41, 45) have suggested relationships between the observed in vivo site-specific glycosylation and the density of the Ser and Thr residues along the PSM polypeptide sequence as quantified by an arbitrarily defined Ser/Thr density function. These relationships were proposed to reflect inhibitory steric effects of neighboring residue glycosylation. Plots of the 70-h ppGalNAc T1 and 250-h ppGalNAc T2 site-specific glycosylation versus the sequence-derived Ser/Thr density are shown in Fig. 3, panels E–H. Consistent with the above studies, an inverse relationship of the extent of glycosylation versus Ser/Thr density is apparent for Ser with ppGalNAc T1 (Fig. 3, panel E) and for both Ser and Thr with ppGalNAc T2 (Fig. 3, panels G and H) (see Fig. 3 legend for statistics). The plot of Thr with ppGalNAc T1 (Fig. 3, panel F) reveals no trends with Ser/Thr density presumably because of the clustering of glycosylation values to a narrow range of very high values.

Modeling Site-specific Glycosylation by Numerical Simulation—To determine whether the observed individual rates of glycosylation indeed reflect the modulation of glycosylation by neighboring residue effects, we performed numerical simulations using a model (described in detail under the “Materials and Methods”) that incorporates the incremental effects of neighboring group glycosylation. The simulation was performed such that the Ser and Thr first-order rate constants, $k_{on}$ and $k_{off}$, were decremented during the course of the simulation as a function of changing neighboring residue glycosylation status, defined as $f_{OG} + OH^-$. The model was formulated to be flexible, permitting independent sensitivities, $W_{OG}$, and $W_{OH}$, to both the presence

4 U. Mandel and H. Clausen (University of Copenhagen), personal communication.
5 The amino acid sequence of the bovine ppGalNAc T1 (47) used in these studies is 99.3% homologous (555 of 559 residues) to the porcine transferase (52), whereas the porcine transferase is 98.7% homologous (552 of 559 residues) to the human transferase (48). To date the sequence of the porcine ppGalNAc T2 homologue has not been reported, although our and human homology would be expected (see Ref. 10). On the basis of the recent work of Schwientek et al. (10), one would expect identical substrate preferences for the homologous enzymes across the species studied in this work.
and absence of neighboring glycosylated Ser and Thr residues over a range of plus and minus 3 residues of the site of glycosylation. Simulations for each transferase were optimized to reproduce the experimental data by manual iterations of the intrinsic rate constants, $k_{\text{Ser}}$ and $k_{\text{Thr}}$, the 12 sequence-specific positional weighting parameters, $W_{\text{OG}}$ and $W_{\text{OH}}$, and the Ser and Thr-specific function, $f(\text{OG}+\text{OH})$, that relates overall neighboring glycosylation status to a fractional rate constant multiplier, as further described under “Materials and Methods.” Goodness of fit was evaluated visually, by least square correlation coefficient, $r^2$, and by standard deviation. For ppGalNAc T1 the four different incubation times give 120 individual glycosylation values for fitting, and the three time points for ppGalNAc T2 provide a set of 85 values. Our goal in developing this model was to test whether the experimentally observed glycosylation time course could be reproduced by the inhibitory effects of neighboring group glycosylation. We therefore have not submitted the model to an exhaustive mathematical minimization procedure, recognizing in particular that the experimental data are subject to significant errors with respect to the net incubation time, overall transferase activity, and measured extent of glycosylation.

Simulation of ppGalNAc T1—By a series of manual iterations it was possible to obtain values of the Ser and Thr rate constants, $k_{\text{Ser}}$ and $k_{\text{Thr}}$, and the positional weighting coefficients, $W_{\text{OG}}$ and $W_{\text{OH}}$, capable of reproducing the experimental site-specific glycosylation data as shown by Figs. 4 and 5 and in Figs. S5 and S6 of the Supplemental Material. As expected, in the absence of neighboring group inhibition (i.e., all $W_{\text{OG}}$ and $W_{\text{OH}}$ values = 0), the simulation gives a uniform extent of Ser and Thr glycosylation as shown in Fig. 4, panel A (for the average), and Fig. S5, panel A, of the Supplemental Material (for the individual time points). Under these conditions relatively poor correlation coefficients, $r^2$, and S.D. for the simulated versus experimental values are obtained ($r^2$ values of 0.383 and 0.55 and S.D. values of 0.308 and 0.235 for Ser and Thr, respectively). The systematic inclusion of $W_{\text{OG}}$ values of 1 as shown in the plots in Fig. 4 and Fig. S5, panels C displays the PSN tandem repeat amino acid sequence.

**Fig. 2.** Glycosylation of the PSN tandem repeat by ppGalNAc T1 and T2. Panels A and B display the glycosylation pattern in sequential order along the tandem repeat for ppGalNAc T1 and ppGalNAc T2, respectively. Increasingly dark gray bars from left to right at each residue represent 5, 15, 35, and 70 h of net incubations with ppGalNAc T1, or 40, 80, and 250 h of net incubations with ppGalNAc T2. The right-most light gray bar for each residue represents the native in vivo glycosylation for each residue as determined previously (45). Data were obtained from the glycosylation data in Table I of the Supplemental Material. Note that the omitted bars in the panels for Thr79 and Ser80 signify the absence of experimental data for these residues. Panel C displays the PSN tandem repeat amino acid sequence.

---

6 The statistical analysis was performed using data obtained from all time points and all residues except for Ser2 and Thr79 which appear to be outliers. As discussed in the text, the glycosylation of Ser2 and Thr79 clearly appears to be affected by additional factors.
A–D, of the Supplemental Material clearly demonstrate an improved fit for both Ser and Thr residues as the effects of neighboring residue glycosylation are incrementally included in the simulation (see Fig. S5 legend for statistical values). Additional changes in the weighting scheme were found to further improve the simulation for both Ser and Thr as shown by panels E and F of the Supplemental Material. Reducing the $W_{OG,3}$ and $W_{OG,2}$ residue weights to 0.5 and the inclusion of a weak free hydroxyamino acid sensitivity, $W_{OH}$, of 0.2 at the +1 position further improved the fit (Fig. 4, panel E), resulting in a noticeable narrowing of the data point scatter for both Ser and Thr residues (compare Fig. S5, panels D and E, of the Supplemental Material). Optimal $r^2$ and S.D. values for Ser ($r^2 = 0.881$, S.D. = 0.068) and Thr ($r^2 = 0.662$, S.D. = 0.140) were obtained by including additional weak hydroxyamino acid sensitivities, $W_{OH}$, at positions −3, +1, and +3 as shown in panel F of Fig. 4 and Fig. S5 of the Supplemental Material.

To confirm that the apparent success of the simulation indeed reflects authentic neighboring group effects and not artifacts of the fitting procedure, we attempted to fit the experimental data to the model after shifting the experimental data...
FIG. 4. Optimization of the simulated glycosylation of the PSM tandem repeat by \( \text{ppGalNAc T1} \) by adjustments of the positional weighting coefficients, \( W_{OG} \) and \( W_{OH} \). The left panels display the values of \( W_{OG} \) and \( W_{OH} \) (black and gray bars, respectively) used for the simulation (see text for their definitions). The right panels show the comparison of the simulated values (black bars) with the experimental values (gray bars), which for ease of presentation represent the combined average glycosylation values for the 5-, 15-, 35-, and 70-h incubation periods. The simulation was performed using Equation 5, and the values for \( C_{1} \text{Ser}, C_{1} \text{Thr}, \) and \( C_{2} \) are given in Equation 4 and displayed in Fig. 1. First-order rate constants, \( k_{\text{Ser}} \) and \( k_{\text{Thr}} \), of 0.022 and 0.09 mol fraction h\(^{-1}\), respectively, were used for all simulations. See Fig. S5 of the Supplemental Material for the statistical analysis of the data.
for each Ser or Thr to the next Ser or Thr residue, respectively, in the tandem repeat. In this manner the experimental data were effectively removed from their original sequence context, and the simulation was expected to fail. This was observed as no set of positional weighting constants could be obtained that were capable of increasing either Ser or Thr $r^2$ above their initial values, obtained in the absence of neighboring group effects, nor could significant improvements in S.D. values be obtained (data not shown).

The full time course of the optimized ppGalNAc T1 simulation is given in Fig. 5, whereas a comparison of the experimental and simulated data at each individual time point (5, 15, 35, and 70 h) is given in Fig. S6 of the Supplemental Material. Considering the inherent experimental errors in the incubation times, transferase activity, and extent of glycosylation, the model reasonably reproduces the site-specific glycosylation for most residues particularly when visualized at each time point as shown in Fig. S6 of the Supplemental Material. An interpretation of the optimized weighting parameters suggests that ppGalNAc T1 is highly inhibited by the glycosylation of neighboring residues plus or minus 3 residues of the site of glycosylation and very weakly inhibited by the presence of unsubstituted hydroxyamino acid residues at the +1 position and perhaps at the −3 and +3 position. The model-derived Ser and Thr first-order rate constants of 0.022 and 0.090 mol fraction/h representing rate constants of 0.38 and 1.6 μmol of GalNAc (mg of ppGalNAc T1) $^{-1}$ h$^{-1}$, respectively, are consistent with previous reports demonstrating that Thr residues are typically significantly more rapidly glycosylated than Ser residues by pp-GalNAc T1.

It is particularly satisfying that the model may provide insight into the origins of the experimentally observed glycosylation behavior. For example, Ser43, which is the most rapidly glycosylated Ser (Fig. 5 and Fig. S6 of the Supplemental Material), is the only Ser with no hydroxyamino acid neighbors within plus or minus 3 residues (see Fig. 2 for the PSM tandem repeat sequence), although as discussed below Ser63 is the only hydroxyamino acid residue preceded by a Pro. Ser77, which has a single Ser at the −3 position, is also found highly glycosylated. Ser62 and Ser63, in the Ser62–64 triad, are very poorly glycosylated due to their proximity to the more rapidly glycosylating Thr60, whereas Ser64 is more readily glycosylated due to its increased distance from Thr60 (see Fig. 4, panel F, and Fig. S6 of the Supplemental Material). Even the "dip" in the glycosylation of the Thr69–70 dyad is predicted by the model due to the inhibitory effects of the glycosylation of neighboring Thr62. The model also predicts that the glycosylation of several residues will likely plateau at values less than 100% (i.e. Ser residues 6, 14, 23, 47, 54, 59, 62, 63, and 80) (see Figs. 2 and 5) and that this is again due to the glycosylation of neighboring Thr or Ser residues. Further evidence of the importance of the inhibitory effects of rapidly glycosylating neighboring Thr residues arises from the high dependence of the fit of the Ser residue glycosylation to the value for the Thr rate constant, i.e. when $k_{\text{Thr}} = 0$, the Ser residue fit considerably worsens ($r^2 = 0.498$, S.D. = 0.161). In contrast, when there is no Ser glycosylation allowed, $k_{\text{Ser}} = 0$, there is essentially no change in the

---

7 The concordance of the predicted plateaus with the experimental data is not always good. This may be due to both errors in the primary experimental data or inaccuracies in the model such as the nature of the arbitrary $f(GG+OH)$ function (Equation 4) that determines the rate constant multiplier based on local glycosylation status. Regardless, the overall correspondence of the experimental data and the predictions of the model consistently show that the rates of glycosylation will be significantly and systematically reduced, solely on the basis of the glycosylation status of neighboring residues.
simulation for the Thr residues using the optimized positional weighting parameter values in Fig. 4, panel F (r² = 0.657 and S.D. = 0.144). These results clearly show that for the majority of the residues in the PSM tandem repeat, the observed site-specific time course of ppGalNAc T1 glycosylation can be explained to a large extent on the basis of the inhibitory effects of the glycosylation status of the neighboring hydroxyamino acid residues.

Two residues, Ser² and Thr⁷⁹, are predicted to be highly glycosylated by the ppGalNAc T1 simulation but are very poorly glycosylated in vitro and in vivo (Figs. 4 and 5 and Fig. S6 of the Supplemental Material). Ser⁴⁰, which flanks these residues, is also poorly glycosylated both by the simulation and experiment. These differences are not due to end effects nor to the inability to cleave these sites by trypsin.²⁸ Interestingly, Ser² is in a sequence nearly homologous to Ser⁴³ (Glu-Thr-Ser-Arg-Ile-Ser²-Val-Ala-Gly-Ser versus Glu-Thr-Ala-Alg-Pro-Ser⁴³-Val-Ala-Gly-Ser) which is found to be the most rapidly glycosylated by GalNAc. In all cases we observed 90% or greater cleavage of the Ser² peptide whereas the Ser⁴³ peptide remains fully soluble. Sec⁵⁷, which flanks these residues, is also poorly glycosylated both by the simulation and experiment. These differences are not due to end effects nor to the inability to cleave these sites by trypsin.²⁸

values of 0.285 and 0.260 for Ser and Thr (data not shown). The plot of the experimental ppGalNAc T2 data versus the optimal simulation for Ser (see Fig. S7, panel E, left panel, of the Supplemental Material) shows considerable scatter. This scatter is shown to be primarily due to differences between the time points as shown by an examination of the plots for the individual time points (Fig. S8 of the Supplemental Material). At each time point the simulation correctly ranks the observed glycosylation for both Ser and Thr. These time point-dependent variations are attributed to our inability to accurately control transferase activity over the lengthy incubation periods utilized in these experiments. As a final validation of the fitting procedure, attempts to fit the experimental data after shifting the data by one residue were performed as for ppGalNAc T1. Again, no consistent set of parameters were found that would improve the correlation coefficients and standard deviations for both residues.

Only a small number of Ser residues are significantly glycosylated in vitro by ppGalNAc T2, i.e. Ser⁴³, Ser⁷, and Ser⁷¹, and these accept GalNAc at rates similar to several Thr residues (Fig. 2, panel B, and Fig. S4, panel B, of the Supplemental Material). This behavior is approximated by the simulation; however, the simulation incorrectly predicts Ser⁴³ to be highly glycosylated (Fig. 6, panel E, Fig. 7, and Fig. S8 of the Supplemental Material). We cannot presently explain the discrepancy for Ser⁴³, but as discussed above for ppGalNAc T1, Ser⁴³ may be poorly glycosylated due to the secondary and/or tertiary structures of the Ser² and Ser⁴³ peptides and on characterizing the specific role of the neighboring residues in each.

Simulation of ppGalNAc T2—By manually adjusting the individual Ser and Thr rate constants, k₉⁰ and k₈₉, and the positional weighting coefficients, W₀⁹ and W₀₉, as was performed for ppGalNAc T1, it was possible to obtain a good simulation for ppGalNAc T2 as shown by Figs. 6 and 7 and by Figs. S7 and S8 of the Supplemental Material. In the absence of neighboring residue effects, (i.e. W₀⁹ and W₀₉ = 0), the r² values are 0.020 and 0.445, whereas the S.D. values are 0.429 and 0.374 for Ser and Thr, respectively (data not shown). The systematic inclusion of W₀⁹ values was found to somewhat improve the fit. For example, W₀⁹ values of 1 from the −2 through +2 positions give r² values of 0.089 and 0.440 and S.D. values of 0.285 and 0.260 for Ser and Thr (data not shown). The additional inclusion of W₀⁹ values of 1 at the −3 and +3 positions further improved the fit for both Ser and Thr (Fig. 6, panel A, and Fig. S7, panel A, of the Supplemental Material, and see the legend of Fig. S7 for statistical values). Only with the inclusion of values for W₀₉, does the simulation significantly improve as shown by Fig. 6 and Fig. S7, panels B–E, of the Supplemental Material. With a W₀₉ value of 0.9, the simulation of Thr⁸⁰ through Thr⁸⁹ significantly improved (Fig. 6, panel B and Fig. S7, panel B, of the Supplemental Material). Setting all W₀₉ values to 0.3 (Fig. 6, panel C, and Fig. S7, panel C, of the Supplemental Material) reproduced the pattern of Thr⁸⁰ through Thr⁸⁹ and improved the simulation for Ser. Limiting the W₀₉ weights to only the −2 through +2 positions was found to reduce the fit giving r² values of 0.222 and 0.463, and S.D. values of 0.185 and 0.163 for Ser and Thr, respectively (data not shown). By increasing the W₀₉, W₀₉ weight to 0.9 while maintaining the remaining W₀₉ values at 0.3, a significant improvement in the simulation for both Ser and Thr was achieved (Fig. 6, panel D, and Fig. S7, panel D, of the Supplemental Material). Further adjustment of the W₀₉ values to 0.5 at positions +3 and −3 resulted in the best overall simulation (Fig. 6, panel E, and Fig. S7, panel E, of the Supplemental Material), giving r² values of 0.532 and 0.778 and S.D. values of 0.116 and 0.092 for Ser and Thr, respectively.

On the basis of the obtained weighting parameters, ppGalNAc T2 appears to be highly sensitive to neighboring glycosylation as well as to the presence of neighboring nonglycosylated hydroxyamino acid residues, especially at the +1 position. The optimized ppGalNAc T2 first-order Ser and Thr rate constants are −4 and 10-fold lower relative to ppGalNAc T1 (0.0055 and 0.008 mol fraction h⁻¹ or 0.094 and 0.14 μmol (mg of ppGalNAc T2)⁻¹ h⁻¹). These values are consistent with the lower activities of ppGalNAc T2 compared with ppGalNAc T1 reported against the same substrates (21, 23, 25). The lower ratio of k₉₀/k₈₉ for ppGalNAc T2 compared with ppGalNAc T1 (1.6 versus 4.9 respectively) also follows the same trends observed previously for these transferases (25).

For example, the best fit for Ser gave values of r² and S.D. of 0.251 and 0.165, respectively, whereas one of the best fits for Thr gave values for r² and S.D. of 0.643 and 0.279, respectively. No set of weighting parameters could be found that simultaneously improved the fitting of both Ser and Thr.

⁴ Previous studies from our laboratory (41, 42, 45) have shown that the observed low glycosylation of Ser², Thr⁷⁹, and Ser⁸⁰ is not an artifact of the inability of trypsin to cleave at Arg⁴ⁱ when these residues are glycosylated by GalNAc. In all cases we observed 90% or greater cleavage at this site as shown by the gel filtration chromatograms in Fig. 2 of both Refs. 41 and 42. The initially reported very high in vitro glycosylation of Ser⁴⁰ (42) has not been confirmed by our more recent studies (41, 45).

⁵ T. A. Gerken and J. Levine, unpublished data.
tertiary structural effects of the peptide. Most of the remaining Ser residues appear experimentally to be refractory to glycosylation by ppGalNAc T2 (Figs. 2, panel B, and 7 and Fig. S4, panel B, of the Supplemental Material). In contrast, the Thr residues appear to be capable of further glycosylation. These observations are reproduced, to the most part, by the ppGalNAc T2 simulation (Fig. 7 and Fig. S8 of the Supplemental Material). We conclude that the very low rates of Ser glycosylation observed for ppGalNAc T2 are best explained in

\[ \text{Simulation Avg} \quad \text{ppGN T2 avg} \]

**FIG. 6.** Optimization of the simulated glycosylation of the PSM tandem repeat by ppGalNAc T2 by adjustments of the positional weighting coefficients, \( W_{OGn} \) and \( W_{OHn} \). The left panels display the values of \( W_{OGn} \) (black) and \( W_{OHn} \) (gray) used for the simulation (see text for their definitions). The right-hand panels show the comparison of the simulated values (black bars) with the experimental values (gray bars), which for ease of presentation represent the combined average glycosylation values for the 40-, 80-, and 250-h incubation periods. The simulation was performed using Equation 5, and the values for \( C_1 \), \( C_2 \), \( C_3 \), and \( C_4 \) given in Equation 4 and displayed in Fig. 1. First-order rate constants, \( k_{Ser} \) and \( k_{Thr} \), of 0.0055 and 0.008 mol fraction h\(^{-1}\), respectively, were used for all simulations. See Fig. S7 of the Supplemental Material for the statistical analysis of the data. Note that no experimental data are available for Thr\(^{79}\) and Ser\(^{80}\).
terms of the inhibitory effects of neighboring nonglycosylated hydroxyamino acid residues and to a lesser extent due to neighboring glycosylation. This contrasts with ppGalNAc T1 whose glycosylation appears to be dominated by the inhibitory effects of neighboring residue glycosylation. Thus, unlike the ppGalNAc T1 simulation, the inhibition of Thr glycosylation, $k_{\text{Thr}} = 0$, does not greatly affect the simulation for Ser ($r^2 = 0.441$, S.D. = 0.130) using the optimized positional weighting parameters of Fig. 6, panel E.

Similar to ppGalNAc T1, the elimination of Ser glycosylation, $k_{\text{Ser}} = 0$, does not affect the simulation of the Thr ($r^2 = 0.773$, S.D. = 0.096).

Of particular interest is the very high sensitivity of ppGalNAc T2 to nonglycosylated Ser and Thr at the $+1$ position and the elevated sensitivities at $-3$ and $+3$ positions. The former would tend to direct glycosylation to the C-terminal residue in hydroxyamino acid residue dyad sequences, as shown for the Ser$^6$ to Ser$^7$ and Thr$^{29}$ to Thr$^{30}$ dyads (see Figs. 2, panel B, 6, panel E, and Fig. S8 of the Supplemental Material). Such preferences are not reported in previous studies (27, 28, 30) on ppGalNAc T2, perhaps due to the limited number of peptides studied and due to the potential for end effects. The overall high sensitivity of ppGalNAc T2 to the neighboring nonglycosylated hydroxyamino acid residue is not presently understood. Perhaps substrate peptide binding is sufficiently weak that the neighboring hydroxyamino acid residues can compete with the site of glycosylation thereby reducing the overall efficiency of GalNAc transfer. Regardless, the values of the positional weighting parameters obtained are useful for comparing and contrasting the properties of the different ppGalNAc transferase.

It is of interest to examine the distribution of Ser and Thr residues neighboring each hydroxyamino acid residue in the PSM tandem repeat to determine whether any given positions are over- or under-represented which could result in the skewing of the positional weighting parameters. From the sequence analysis (see Table 2 of the Supplemental Material) the distribution of neighboring hydroxyamino acid residues is remarkably uniform for both Ser and Thr, with each position having between 9 and 11 hydroxyamino acid residues. We conclude that an uneven distribution of hydroxyamino acid residues is not responsible for the weighting parameters obtained for either transferase.

**DISCUSSION**

Studies were undertaken to examine the activities of purified ppGalNAc transferases T1 and T2 against the porcine submaxillary mucin tandem repeat substrate whose *in vivo* glycosylation pattern has been extensively characterized (41, 42, 45). It was of interest to determine the extent that the single transferases could reproduce the observed *in vivo* glycosylation pattern and to determine whether the inverse correlations of glycosylation by GalNAc with hydroxyamino acid density and presumably glycosylation density would be reproduced *in vitro*. These studies have successfully addressed both issues, suggesting that ppGalNAc T1 is a major contributor to the mucin’s glycosylation and confirming, by a kinetic modeling approach, that local glycosylation status can account for much of the glycosylation behavior of both transferases.

Experimental evidence for neighboring glycosylation decreasing the rate of glycosylation has been reported previously (26, 27, 30) for ppGalNAc T1 through T4 against a range of small glycopeptides substrates. However, these studies did not yield predictive rules, perhaps due to the presence of end effects due to the use of relatively short (5–25 residues) glycopeptide substrates. The present work was performed on the relatively intact oligomeric tandem repeat domains derived from PSM; therefore, end effects should be absent in the final analysis of the glycosylation of the isolated tandem repeat. An
A Kinetic Model of Mucin O-Glycosylation

The additional advantage of our approach is that the rate data for all of the Ser and Thr residues in the PSM tandem repeat are obtained under exactly identical conditions, thereby eliminating additional experimental variables that may interfere with their comparison.

The demonstration that the time course of glycosylation of the PSM tandem repeat can be nearly completely accounted for on the basis of neighboring group glycosylation status, for both ppGalNAc T1 and T2, is particularly interesting as these results seem to contradict the many previous studies demonstrating that the ppGalNAc transferases possess clear substrate preferences related to peptide sequence and composition. Many in vivo and in vitro studies (26, 38–40) have demonstrated the importance of neighboring Pro residues and the modulating effects of charged residues. Statistical data base studies on O-glycosylation also demonstrate a high prevalence for Pro, Ser, and Thr residues neighboring the sites of O-glycosylation (33–35). In addition, several ppGalNAc transferase isoforms have sufficiently different substrate specificities, such that unique transferase-specific peptide acceptors have been identified (13, 21). On the basis of our prior knowledge the only feasible explanation for the success of our modeling is that the PSM tandem repeat, having evolved to be efficiently O-glycosylated, is composed of exceptionally good acceptor sites having nearly identical initial rates of O-glycosylation (indeed as implemented in the model). Only under such conditions would the inhibitory effects of neighboring glycosylation be readily revealed. As we have discussed above, the observation that Ser
d⁰ and perhaps Thr⁷⁹ are far less glycosylated than expected by our kinetic model suggests that these residues may be intrinsically poor acceptors. It is therefore anticipated that additional refinements in the model, taking into account rate constant decreases or enhancements due to specific neighboring residues or peptide sequences, may be required before the glycosylation of other mucin peptide domains can be successfully modeled by this approach.

It should be noted that the inclusion into our model of the inhibitory effects of neighboring nonglycosylated hydroxymino acid residues in fact introduces a sequence-specific component, effectively recognizing that ppGalNAc transferases may possess unfavorable hydroxymino acid sequence motifs. On the basis of the modeling, ppGalNAc T2 clearly shows a highly specific sensitivity to the presence of neighboring hydroxymino acid residues, particularly at the +1 position, which is only weakly observed for ppGalNAc T1. Whether other ppGalNAc transferases exhibit similar sensitivities remains to be determined. Although the above may appear to conflict with the statistical analysis of O-glycosylation sites which suggests the presence of Ser and Thr as predictors of O-glycosylation (33–35), it has been shown that the association of Ser and Thr with O-glycosylation is the result of the statistical clustering of Ser and Thr residues (34).

In summary, a kinetic model based on first principles has been described as capable of reproducing the in vitro glycosylation of the PSM tandem repeat domain by both ppGalNAc T1 and T2. Key to the model is the reduction of the rate constant proportional to the neighboring residue glycosylation status. An analysis of the positional sequence weighting coefficients reveals that both ppGalNAc T1 and T2 possess sensitivities to the neighboring glycosylation state up to plus and minus 3 residues of the site of glycosylation, generally in keeping with previous O-glycosylation site analysis (33–35). Each transferase has been found to possess common and unique sensitivities, with ppGalNAc T2 showing significantly higher sensitivity to neighboring nonglycosylated hydroxymino acid residues than ppGalNAc T1. These findings support our previous in vivo studies that revealed an inverse relationship with the extent of glycosylation and Ser/Thr density (45). For those cases where the model predicts greater glycosylation than experiment, we propose as a plausible explanation the presence of intrinsically poor acceptor substrates. This work demonstrates that in addition to the intrinsic propensity of a substrate for O-glycosylation dictated by peptide sequence and conformation, the glycosylation states of neighboring residues play equally important roles in determining mucin O-glycosylation.

Acknowledgments—We thank Marc Gilmore for technical assistance and Drs. Himan Sternlicht, Vernon Anderson, Eckard Jankowsky, and Frank Sonnichsen for the helpful discussions.

REFERENCES

1. Perez-Vilar, J., and Hill, R. L. (1999) J. Biol. Chem. 274, 31751–31754
2. Van den, S. P., Rudd, P. M., Dwek, R. A., and Opdenakker, G. (1998) Crit. Rev. Biochem. Mol. Biol. 33, 151–208
3. Fukuda, M., and Tsuibi, S. (1999) Biochem. Biophys. Acta 1455, 205–217
4. Berger, E. G. (1999) Biochem. Biophys. Acta 1455, 255–268
5. Rudd, P. M., Wormald, M. R., Stanfield, R. L., Huang, M., Mattson, N., Spear, J. A., DiGennaro, J. A., Row, J. S., Dwek, R. A., and Wilsen, I. A. (1999) J. Biol. Chem. 274, 351–366
6. Taylor–Papadimitriou, J., Burchell, J., Miles, D. W., and Dalziel, M. (1999) Biochim. Biophys. Acta 1455, 301–313
7. Taylor–Papadimitriou, J., Burchell, J., Miles, D. W., and Dalziel, M. (1999) Biochim. Biophys. Acta 1455, 301–313
8. Brunak, S., Jorgensen, T., and Brunak, S. (1998) Glycoconjug. J. 15, 115–130
9. Chou, K. C. (1995) Protein Sci. 4, 1365–1383

49861
36. Yoshida, A., Suzuki, M., Ikenaga, H., and Takeuchi, M. (1997) *J. Biol. Chem.* **272**, 16884–16888
37. Nehrke, K., Hagen, F. K., and Tabak, L. A. (1996) *J. Biol. Chem.* **271**, 7061–7065
38. Nehrke, K., Ten Hagen, K. G., Hagen, F. K., and Tabak, L. A. (1997) *Glycobiology* **7**, 1053–1060
39. Hennébicq, S., Tetaert, D., Soudan, B., Briand, G., Richet, C., Gagnon, J., and Degand, P. (1998) *Glycoconj. J.* **15**, 275–282
40. Sugahara, T., Pixley, M. R., Fares, F., and Boersma, A. (1996) *J. Biol. Chem.* **271**, 20797–20804
41. Gerken, T. A., Owens, C. L., and Pasumarthy, M. (1998) *J. Biol. Chem.* **273**, 26580–26588
42. Gerken, T. A., Owens, C. L., and Pasumarthy, M. (1997) *J. Biol. Chem.* **272**, 9709–9719
43. Timpte, C. S., Eckhardt, A. E., Abernethy, J. L., and Hill, R. L. (1988) *J. Biol. Chem.* **263**, 1081–1088
44. Eckhardt, A. E., Timpte, C. S., DeLuca, A. W., and Hill, R. L. (1997) *J. Biol. Chem.* **272**, 35324–35329
45. Gerken, T. A., Gilmore, M., and Zhang, J. (2002) *J. Biol. Chem.* **277**, 7736–7751
46. Gerken, T. A., Gupta, R., and Jentoft, N. (1992) *Biochemistry* **31**, 639–648
47. Homa, F. L., Hollander, T., Lehman, D. J., Thomsen, D. R., and Elhammer, A. P. (1995) *J. Biol. Chem.* **268**, 12609–12616
48. White, T., Bennett, E. P., Takio, K., Sorensen, T., Bonding, N., and Clausen, H. (1995) *J. Biol. Chem.* **270**, 24156–24165
49. Homa, F. L., Baker, C. A., Thomsen, D. R., and Elhammer, A. P. (1995) *Protein Expression Purif.* **6**, 141–148
50. Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J. (1951) *J. Biol. Chem.* **193**, 265–275
51. Mandel, U., Hasson, H., Therkildsen, M. H., Rygaard, J., Jakobsen, M. H., Juhl, B. R., Dabelsteen, E., and Clausen, H. (1999) *Glycobiology* **9**, 43–52
52. Yoshida, A., Hara, T., Ikenaga, H., and Takeuchi, M. (1995) *Glycoconj. J.* **12**, 824–828
53. Wragg, S., Hagen, F. K., and Tabak, L. A. (1995) *J. Biol. Chem.* **270**, 16947–16954
54. Gerken, T. A., and Dearborn, D. G. (1984) *Biochemistry* **23**, 1485–1497