Database Analysis of *O*-Glycosylation Sites in Proteins

T. Hema Thanka Christlet and K. Veluraja
Department of Physics, Manonmaniam Sundaranar University, Tirunelveli 627 012, Tamil Nadu, India

ABSTRACT Statistical analysis was carried out to study the sequential aspects of amino acids around the *O*-glycosylated Ser/Thr. 992 sequences containing *O*-glycosylated Ser/Thr were selected from the O-GLYCBASE database of *O*-glycosylated proteins. The frequency of occurrence of amino acid residues around the glycosylated Ser/Thr revealed that there is an increased number of proline residues around the *O*-glycosylation sites in comparison with the nonglycosylated serine and threonine residues. The deviation parameter calculated as a measure of preferential and nonpreferential occurrence of amino acid residues around the glycosylation site shows that Pro has the maximum preference around the *O*-glycosylation site. Pro at +3 and/or −1 positions strongly favors glycosylation irrespective of single and multiple glycosylation sites. In addition, serine and threonine are preferred around the multiple glycosylation sites due to the effect of clusters of closely spaced glycosylated Ser/Thr. The preference of amino acids around the sites of mucin-type glycosylation is found likely to be similar to that of the *O*-glycosylation sites when taken together, but the acidic amino acids are more preferred around Ser/Thr in mucin-type glycosylation when compared totally. Aromatic amino acids hinder *O*-glycosylation in contrast to *N*-glycosylation. Cysteine and amino acids with bulky side chains inhibit *O*-glycosylation. The preference of certain potential sequence motifs of glycosylation has been discussed.

INTRODUCTION

Glycosylation is a common and highly diverse co- and post-translational protein modification reaction, which occurs in most eukaryotic cells. Such modifications can be divided broadly into two categories, *N*-linked glycosylation and *O*-linked glycosylation. In *N*-linked glycosylation the carbohydrate moiety is attached to the amide nitrogen of the side chain of asparagine, when asparagine is part of the consensus sequence Asn-X-Ser/Thr (Marshall, 1972). This consensus sequence is a necessary but not a sufficient condition for *N*-glycosylation to occur. A proline residue at X position prevents *N*-glycosylation (Bause and Hettkamp, 1979; Bause, 1983) and a proline residue in the sequence beyond Ser/Thr inhibits *N*-glycosylation (Gavel and von Heijne, 1990).

In *O*-linked glycosylation the carbohydrate moiety is covalently linked to the hydroxyl oxygen of the hydroxy-amino acids serine and threonine. In addition, *O*-glycosylation also occurs as a primary modification of tyrosine and as a secondary modification of 5-hydroxylysine and 4-hydroxyproline. *O*-linked glycosylation serves a variety of functions such as ligands for selectins and resistance to proteolysis of stem regions of membrane proteins (Jentoft, 1990; Hart, 1992) and are involved in recognition phenomena (Fukuda, 1991). Seven different classes of *O*-linked glycosylation were identified: mucin type (Strous and Dekker, 1992; Carraway and Hull, 1991), intracellular type (Haltiwanger et al., 1992), xyloglycan type (Yanagishita and Hascall, 1992), collagen type (Spiro, 1973), clotting factor type (Nishimura et al., 1992), fungal type (Hausler et al., 1992), and plant type (Allen et al., 1978). To date, a consensus primary amino acid sequence for *O*-glycosylation has not been identified. Different authors have proposed different structural motifs for *O*-glycosylation (Young et al., 1979; Muller et al., 1997; Yoshida et al., 1997; Gooley et al., 1991). Lehle and Bause (1984) postulated that accessibility of potential *O*-glycosylation sites rather than a specific sequence may be a prerequisite for *O*-glycosylation.

The primary, secondary, tertiary, and quaternary structures are important for efficient *O*-glycosylation (Aubert et al., 1976; Fiat et al., 1980; Dahms and Hart, 1986; Wang et al., 1993). It has been observed that in contrast to *N*-glycosylation, proline residues are commonly found in the sequence close to the *O*-glycosylated serine or threonine (Young et al., 1979). Several predictive methods are available for estimating the relative propensity for a given Ser or Thr to be glycosylated (Elhammer et al., 1993; Hansen et al., 1995; Chou, 1995). Statistical studies by Wilson and his co-workers (1991) have shown that proline has a high frequency of occurrence at +3 and −1 positions relative to the glycosylated Ser/Thr. It has also been proposed that proline residues from −4 to +4 positions are important for *O*-glycosylation (Elhammer et al., 1993). A study on recombinant erythropoeitin (rEPO) shows that proline at −1 and +1 sites enhances *O*-glycosylation (Elliott et al., 1994). In vitro experiments done by Elhammer et al. (1993) using bovine colostrum transferase and matrix statistics revealed that not only proline, but also serine and threonine at all positions from -4 to +4 favored *O*-glycosylation. Studies by O’Connell and co-workers (1991) revealed that a proline, alanine, serine, or threonine at +3, −1, −6, and −3 positions was associated with glycosylation whereas a charged residue at these positions was associated with nonglycosylation. Thus, the above theoretical and experimental investigations indicate that the proline from −6 to +4 positions

Received for publication 28 July 2000 and in final form 30 October 2000.

Address reprint requests to Dr. K. Veluraja, Department of Physics, Manonmaniam Sundaranar University, Tirunelveli 627 012, Tamil Nadu, India. Tel.: 0462-333887; Fax: 0462-322973; E-mail: bio@md3.vsnl.net.in.

0006-3495/01/02/952/09 $2.00

plays a vital role in the process of glycosylating the proteins. Studies on the porcine submaxillary mucin suggest that sequences with Gly at positions -2 and $+2$ are associated with higher degree of glycosylation whereas Gly at -3 and $+3$ positions may be associated with reduced glycosylation (Gerken et al., 1997). Many authors suggested the importance of having amino acids with small side chains at the $+2$ and -2 positions (Elhammer et al., 1993; Hansen et al., 1995; O'Connell et al., 1991).

The sequences of a number of *O*-glycosylated proteins have been published in recent years. SWISSPROT, PIR, PROSITE, PDB, EMBL, HSSP, LISTA, and MIM databases contain *O*-glycosylated entries and O-GLYCBASE is an updated database of information on glycoproteins and their *O*-linked glycosylation sites (Gupta et al., 1999). Here we present the results of an attempt, based on an analysis of amino acid sequence around the experimentally predicted *O*-glycosylated sites from the O-GLYCBASE database. The positional preference of amino acids around the site of glycosylation has been investigated using statistical methods.

MATERIALS AND METHODS

O-GLYCBASE is a revised database of *O*-glycosylated proteins (Hansen et al., 1997, 1998; Gupta et al., 1999). Version 4.03 has 180 glycoprotein entries. The database is nonredundant in the sense that it contains no identical sequences, unless there are conflicting glycosylation data. The source of these glycoproteins is mainly mammals, fish, birds, fungi, and plants. In this database the sugars that are involved in linking the carbohydrate moiety to Ser/Thr residues of the protein are GalNAc, Xyl, GlcNAc, Glc, Fuc, Man, and Gal. Version 4.03 of this database is accessible via the Internet.

In this version of the O-GLYCBASE database, there are a total of 992 experimentally verified *O*-glycosylation sites, of which 339 are Ser and 653 are Thr sites. These 992 glycosylated sites were selected for the present investigation. A sequence sub-database was prepared by considering the central residue as the glycosylated Ser/Thr with a selection of 10 amino acid residues on either side of it ($i - 10$ to $i + 10$; i is the position of glycosylated Ser/Thr). Thus, the window size of the amino acid segments selected for our analysis is 21.

The observed frequency of occurrence of amino acid residues at each position in the selected window size is calculated as

$$F_{\mathrm{observed}}(\mathrm{X})_{\mathrm{P}} = \frac{\sum N_{\mathrm{observed}}(\mathrm{X})_{\mathrm{P}}}{m},$$

where $N_{\mathrm{observed}}(\mathrm{X})_{\mathrm{P}}$ is the observed count for the amino acid X at position P and m is the total number of glycosylated sequence segments. The expected frequency of occurrence $F_{\mathrm{expected}}(\mathrm{X})$ and the expected count $N_{\mathrm{expected}}(\mathrm{X})$ are given by

$$F_{\mathrm{expected}}(\mathrm{X}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} N_{\mathrm{i}}(\mathrm{X})}{\sum_{\mathrm{i}=1}^{\mathrm{n}} T_{\mathrm{i}}},$$

$$N_{\mathrm{expected}}(\mathrm{X}) = mF_{\mathrm{expected}}(\mathrm{X}),$$

where $N_{\mathrm{i}}(\mathrm{X})$ is the number of amino acid residues of type X in protein i, T_{i} is the total number of amino acids in protein i, and n is the total number of proteins (in this case, 180).

To estimate the preferential occurrence of amino acid residues at particular positions around the glycosylated site, a quantity called deviation parameter (DP) was calculated. This parameter is the normalized factor of the difference between observed and expected frequencies and is expressed as a percentage. The deviation parameter at position P for the amino acid X is calculated as

$$\mathrm{DP}(\mathrm{X})_{\mathrm{P}} = \frac{F_{\mathrm{observed}}(\mathrm{X})_{\mathrm{P}} - F_{\mathrm{expected}}(\mathrm{X})}{F_{\mathrm{expected}}(\mathrm{X})} \times 100.$$

A positive $\mathrm{DP}(\mathrm{X})_{\mathrm{P}}$ indicates positive preference for the type X amino acid at position P and a negative $\mathrm{DP}(\mathrm{X})_{\mathrm{P}}$ indicates negative preference for the type X amino acid at position P. The statistical significance of the given DP values depends upon the difference between the observed and expected counts ($N_{\mathrm{observed}}(\mathrm{X})_{\mathrm{P}} \sim N_{\mathrm{expected}}(\mathrm{X})$) and the σ values, where

$$\sigma = \sqrt{N_{\mathrm{observed}}(\mathrm{X})_{\mathrm{P}}}.$$

If the difference between the observed and expected counts ($N_{\mathrm{observed}} \sim N_{\mathrm{expected}}$) is $\geq 2\sigma$ at a particular position, the DP value at that position for that amino acid is considered to be statistically significant.

The data set containing 992 glycosylation sites was subdivided into two data sets with single glycosylation sites and multiple glycosylation sites, because some proteins are *O*-glycosylated at one or a few isolated sites whereas other proteins have clusters of closely spaced *O*-glycosylation sites within the chosen window size. DP and σ values were computed for each of the 20 amino acids at each of the positions in the window size selected for the sequences in these two groups. In addition, the computations were also carried out for the mucin-type glycosylation.

The unmodified serine and threonine residues of the same protein were used to form a data set of nonglycosylated sequences. Nonglycosylated Ser/Thr were selected based on the criterion that glycosylated Ser/Thr should not occur within the chosen window size of 21 residues. For a comparative study, the parameters computed for the glycosylated sites were repeated for the nonglycosylated sites. To throw more light on the positional preference of amino acids, the window size was varied from 7 to 21 amino acids in which the central amino acid is the glycosylated Ser/Thr.

RESULTS AND DISCUSSION

The database consists of 180 glycoproteins containing a total of 992 *O*-glycosylated sites. These glycoproteins also contain 8952 nonglycosylated serine and threonine residues that are used for the comparative study, based on the selection mentioned in Materials and Methods. Based on the sugar unit attached to Ser/Thr, the seven types of *O*-glycosylation were separated out. In this database of 180 proteins, in 97 proteins the sugar linked is GalNAc (mucin type, 644 sites), in 24 proteins the sugar linked is GlcNAc (intracellular type, 87 sites), in 9 proteins it is Gal-Gal-Xyl attached to Ser (xyloglycan type, 15 sites), in 11 proteins it is Man (fungal type, 158 sites), in 13 proteins it is either Fuc or Xyl-Glc or Xyl-Xyl-Glc (clotting factor type, 24 sites), in 1 protein it is Gal (plant type, 1 site), and in 25 proteins the sugar linked is not known or unspecified or sialic acid. Except mucin type, the other types of glycosylation suffer from a lack of a sufficient number of glycosylation sites to have reliable statistics for interpretation, and hence the calculations are restricted to the whole database (992 glycosylation sites) as such and to the mucin-type glycosylation (644 glycosylation sites).

Frequency of occurrence of amino acids in the whole database

The frequency of occurrence of each of the 20 amino acids at each position around the *O*-glycosylated Ser/Thr was computed. The frequency of amino acid residues differs around glycosylated and nonglycosylated sites. There is a high frequency of occurrence of proline, serine, threonine, and alanine residues around the glycosylation sites (Fig. 1 *A*). In addition to this, glycine and valine residues also possess a high frequency of occurrence around the glycosylation sites. On the other hand, in the data set containing nonglycosylated sequence segments, the frequency of occurrence of proline, serine, and threonine residues are moderate (Fig. 1 *B*).

Preference of amino acids around multiple and single glycosylation sites

Multiple glycosylation sites contain clusters of closely spaced *O*-glycosylation sites; i.e., adjacent and consecutive serine and threonine residues are glycosylated and are found in 98 glycoproteins of our data set. DP and σ values were calculated, and those that have statistically significant DP values are shown by an asterisk in Table 1. Proline, serine, threonine, and alanine are highly preferred around the glycosylated Ser and Thr of this group. The DP value of proline is maximal at the $+3$ position (262) followed by the value at the $+7$ position (165) and is minimal at the -8 position (-9). Proline is preferred from positions -9 to $+9$ (except at ± 8, ± 4, and -2) around multiple glycosylation sites. The amino acids serine and threonine have significant DP values due to the fact that the serine and threonine within the window size is also glycosylated in the multiple glycosylation sequences. Gly, Ala, and Val are preferred, implying that amino acids with small side chains are more favored around multiple glycosylation sites. In addition to this the amino acid Asp is preferred at positions -8, -3, and $+10$; His at positions -3 and $+5$; and Arg at position $+6$, as indicated by the positive significant DP values given in Table 1.

The same calculations were repeated for the isolated sites. DP and their corresponding σ values revealed that in this group also proline possesses statistically significant positive DP values (Table 2). The DP value of proline varies from -16 to 354 having the maximum at the $+3$ position (354) followed by its value at the -1 position (280). Proline is preferred at positions from -3 to $+5$ around isolated sites and valine is preferred at the -1 position. All other results highlight the fact that proline is preferred near to the site of glycosylation irrespective of whether the site is singly glycosylated or multiply glycosylated. It is an interesting feature that Ser and Thr are preferred only around multiple glycosylation sites due to the presence of clusters of glycosylated Ser and Thr, but Ser and Thr are less preferred around the single glycosylation sites (Table 2). On the other hand, proline is preferred close to the glycosylated Ser/Thr and it is highly preferred at the -3, -1, $+1$, $+2$, $+3$, and $+5$ positions in both cases. Also valine is preferred at the -1 position in these two groups.

All other amino acids show negative DP values or insignificant positive DP values for the multiple and single glycosylation sequences. Cys possesses statistically significant negative DP values (Table 1 and 2). The ability of Cys to form disulfide bonds may hinder glycosylation and thus explain the low frequency of occurrence of Cys residues close to the glycosylation site. The aromatic amino acids Phe, Trp, and Tyr possess statistically significant negative DP values implying aromatic amino acids are less preferred near the site of *O*-glycosylation. Lys, Leu, Gln, and Asn are also selected against at various positions close to the *O*-glycosylation site.

Positional preference of proline

As mentioned earlier, proline is largely preferred around the *O*-glycosylation sites. To find out up to what position the proline is preferred, the DP values were calculated for various window sizes ranging from 7 to 21 excluding the glycosylated Ser/Thr within the window size selected and also three amino acids from the ends. DP values for serine and threonine glycosylation were calculated separately. The same calculations were repeated for the sequences of the nonglycosylated data set, and the results are plotted in Fig. 2, *A–D*. From the figures it is seen that presence of proline usually enhances threonine glycosylation rather than serine glycosylation. Although proline favors Thr glycosylation, the presence of proline at -2 and $+2$ positions favors Ser glycosylation rather than Thr glycosylation irrespective of the window size. Also there are a pronounced number of proline residues at -1 and $+3$ positions, which is not altered even though the window size is changed. They show smaller DP values when positioned farther away from the glycosylation site and the preference of proline is reduced beyond ± 3. This shows that proline may not have pronounced effect beyond ± 3. There is a general notion that proline at $+3$ and -1 positions enhances glycosylation (O'Connell et al., 1991, 1992; Elliott et al., 1994; Wilson et al., 1991; Hansen et al., 1995), which is substantiated by deviation parameter calculations. On the other hand, proline is not preferred around the nonglycosylated Ser/Thr (Fig. 2), which further provides evidence for the high significance of proline near Ser/Thr in the *O*-glycosylation process.

Significance of sequence motifs in *O*-glycosylation

Different types of sequence motifs were reported by various authors (Yoshida et al., 1997; Young et al., 1979; Gooley et al., 1991; Pisano et al., 1993) based on experimental studies and in vitro and in vivo analyses. Most of the sequence motifs

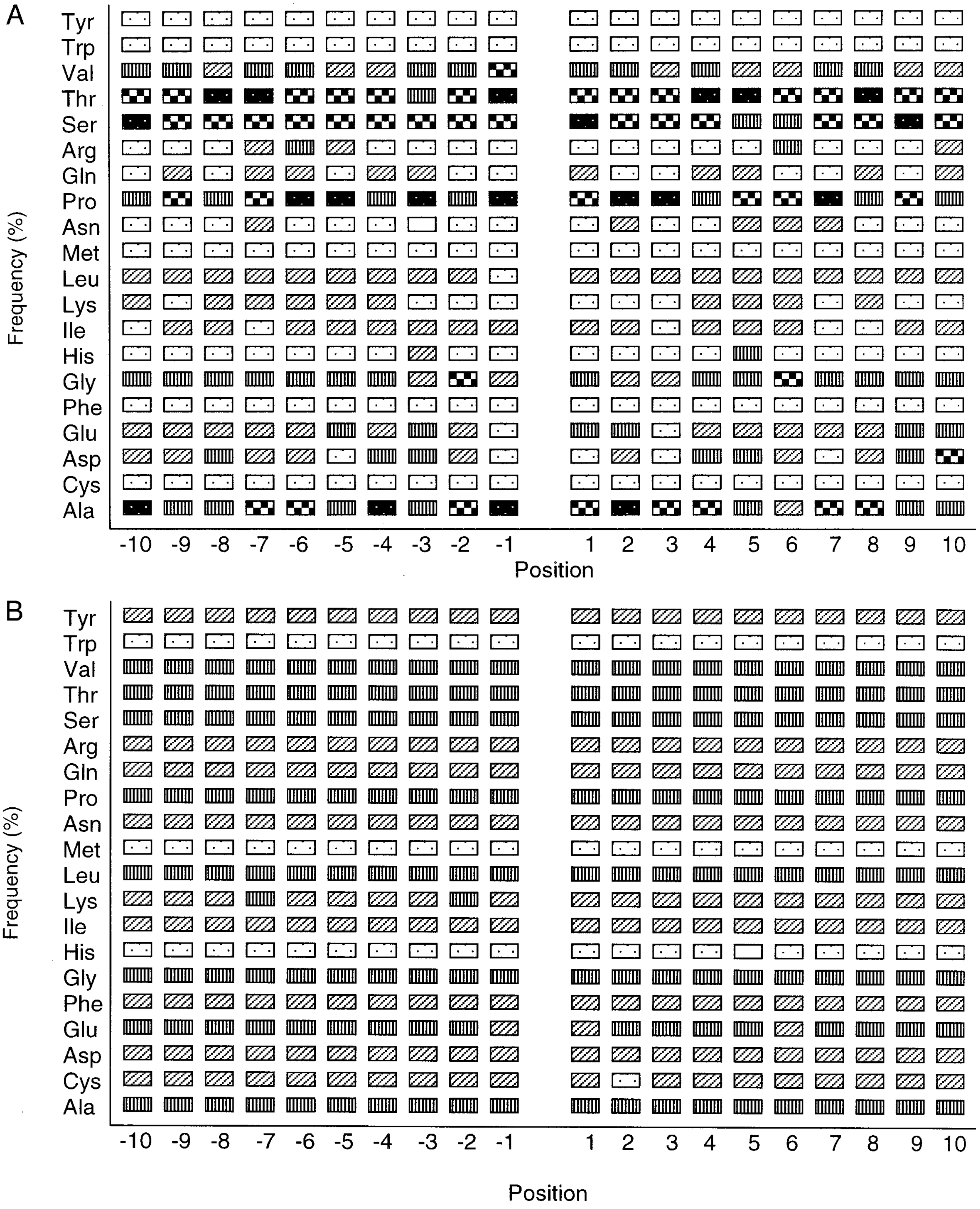

FIGURE 1 Frequency of occurrence of amino acids around glycosylated Ser/Thr (*A*) and nonglycosylated Ser/Thr (*B*). The frequency of each amino acid is expressed as a percentage in this plot for positions −10 to +10. The shading at each location indicates the percentage frequency of a residue: ■, >15; ▦, 10–15; ▥, 5–10; ▨, 2–5; ⬚, <2.

TABLE 1 Deviation parameters for the amino acids at positions from −10 to +10 for the multiple glycosylation sequences using the whole database

Amino	DP values at positions																			
acid	$i - 10$	$i - 9$	$i - 8$	$i - 7$	$i - 6$	$i - 5$	$i - 4$	$i - 3$	$i - 2$	$i - 1$	$i + 1$	$i + 2$	$i + 3$	$i + 4$	$i + 5$	$i + 6$	$i + 7$	$i + 8$	$i + 9$	$i + 10$
Ala	173*	41*	−11	47*	55*	−1	143*	−11	93*	106*	61*	134*	99*	94*	−12	−30*	42*	49*	−7	30
Cys	−47*	−55*	−68*	−64*	−88*	−52*	−88*	−88*	−61*	−88*	−65*	−81*	−53*	−73*	−69*	−65*	−73*	−73*	−34	−65*
Asp	−44*	−21	116*	4	−24	−63*	41	96*	7	−82*	−66*	−46*	−73*	42	30	17	−60*	−51*	28	127*
Glu	−52*	−41*	−40*	−46*	−49*	−19	−44*	32	−31*	−73*	22	11	−65*	−43*	−28	−29*	−51*	−29*	−1	17
Phe	−62*	−47*	−43*	−74*	−62*	−74*	−63*	−59*	−67*	−63*	−78*	−60*	−60*	−63*	−67*	−52*	−56*	−67*	−45*	−56*
Gly	26	3	18	38*	−30*	24	12	−43*	110*	−37*	39*	−39*	−43*	18	15	74*	−8	22	41*	27
His	−40	−30	−41	−26	−66*	−61*	−27	175*	−42*	−47*	−71*	−43*	−67*	−28	182*	−9	0	−62*	−52*	−38
Ile	−38*	−13	−8	−36*	−27	−27	−9	−25	8	−22	−28	−37*	−40*	−34	−34	−31	−43*	−43*	−34	−37*
Lys	−40*	−65*	−31	−48*	−46*	−49*	−44*	−59*	−66*	−73*	−64*	−59*	−78*	−54*	−49*	−45*	−57*	−40*	−83*	−61*
Leu	−53*	−49*	−59*	−71*	−68*	−49*	−53*	−59*	−63*	−74*	−69*	−65*	−65*	−57*	−46*	−47*	−58*	−61*	−63*	−65*
Met	−55*	−33	−63*	−12	−5	−13	−13	1	−21	−71*	−22	−22	−29	−50*	−14	−64*	−43	−14	−57*	−57*
Asn	−52*	−56*	−38*	−25	−50*	−41*	−45*	−33	−36*	−51*	−52*	−33	−52*	−42*	−36*	−33	−27	−36*	−36*	−45*
Pro	−6	97*	−9	58*	112*	161*	4	113*	23	93*	55*	107*	262*	22	41*	61*	165*	−6	86*	7
Gln	−47*	−44*	−45*	−42*	−40*	−56*	−51*	−19	−43*	−46*	−33	−65*	−41*	−33	−28	−57*	−49*	−52*	−46*	−46*
Arg	−59*	−52*	−77*	−47*	40	−45*	−58*	−75*	−78*	−58*	−66*	−53*	−80*	−68*	−46*	64*	−56*	−68*	−51*	−34*
Ser	93*	69*	37*	36*	62*	54*	52*	31	32	84*	122*	43*	67*	29	7	13	74*	71*	116*	60*
Thr	73*	98*	204*	131*	91*	95*	83	31	58*	109*	58*	78*	46*	145*	154*	100*	32	203*	40*	78*
Val	−19	−16	−20	−14	−10	−20	−27	−8	−12	90*	−1	−13	−21	−13	−16	−33*	48*	−16	−26	−29*
Trp	−61*	−22	−90*	−52	−4	−81*	−53	−53	−72*	−81*	−81*	−63*	−63*	−63*	−81*	−53*	−53*	−72*	−72*	−63*
Tyr	−71*	−63*	−55*	−42*	−67*	−43*	−59*	−47*	−48*	−52*	−68*	−40*	−40*	−72*	−72*	−24	−60*	−8	−28	−32

*Statistically significant DP values.

contain a proline at the +3 position. In our data set also, certain sequence motifs are present repeatedly with proline at the +3 position. From all the computations discussed earlier, it is interesting to note that for both single and multiple glycosylation sequence segments, the presence of proline at the −1 position and/or at the +3 position is statistically significant: 29% of the total glycosylated sequences (992) contain proline at the +3 position, 16% contain proline at the −1 position, and 9% contain proline at both −1 and +3 positions. A total of 287 sequences carry proline at the +3 position, of which 16% of

TABLE 2 Deviation parameters for the amino acids at positions from −10 to +10 for the single glycosylation sequences using the whole database

Amino	DP values at positions																			
acid	$i - 10$	$i - 9$	$i - 8$	$i - 7$	$i - 6$	$i - 5$	$i - 4$	$i - 3$	$i - 2$	$i - 1$	$i + 1$	$i + 2$	$i + 3$	$i + 4$	$i + 5$	$i + 6$	$i + 7$	$i + 8$	$i + 9$	$i + 10$
Ala	−59*	1	20	−50	8	−11	−12	16	73	41	70	41	−14	23	4	−24	−14	−33	−14	−52
Cys	24	−25	−26	46	21	93	0	0	65	−76*	16	0	62	−30	62	16	39	16	201	16
Asp	17	75	29	−13	0	0	12	68	−58	−18	−45	−31	−31	9	36	−18	−18	50	9	36
Glu	2	2	12	−10	77	22	97	−12	8	−89*	27	−68*	−68*	−25	−25	−4	−36	−57*	−14	6
Phe	13	13	−77*	−11	−11	−55	−13	−56	−56	−36	−15	−57	−15	−57	−36	−15	132	−78*	−15	5
Gly	7	27	35	15	42	62	31	−43	21	37	18	55	−35	−54*	−17	−8	92	−45	46	18
His	−69	−39	19	−11	−41	−70	−70	−41	0	0	−43	−15	−43	−43	−15	12	−43	−71*	−15	125
Ile	−22	94	−61	−42	−5	−24	−62	30	85	−81*	8	45	−81*	−27	−9	−9	45	−45	8	26
Lys	16	1	29	28	−71*	0	−30	−44	−72*	−72*	−86*	−18	−72*	22	−4	8	−4	8	22	−59*
Leu	85	−7	−16	−33	−34	14	−2	21	−19	−76*	−44	−36	−44	18	−29	−21	−13	−29	−13	−44
Met	−8	−8	−55	−10	77	0	−56	0	0	0	0	0	−15	27	69	−15	−15	69	−15	27
Asn	35	−22	−42	13	−62	−24	−62	−7	−7	0	−63*	−27	−27	−27	134	26	−45	−27	−9	8
Pro	39	0	7	−12	64	−12	62	147*	89	280*	122*	122*	354*	141*	131*	48	48	66	−16	39
Gln	−31	19	1	−49	−16	16	96	47	14	27	43	27	−36	75	11	11	−20	27	−4	43
Arg	−22	−22	22	82	−9	50	4	−70*	18	−85*	15	−27	−27	−71*	−42	1	30	15	15	−27
Ser	46	3	−48	60	−7	−32	40	−33	−17	4	28	36	20	52	−11	−3	−27	68	−35	−3
Thr	−79*	−69*	0	28	−11	−11	−51	−13	−32	22	−34	−24	−34	−34	−52*	−5	−24	22	−43	−15
Val	1	−32	−10	−11	−45	−23	−45	40	18	131*	−5	36	−26	−26	−57*	15	−57*	36	26	26
Trp	0	63	60	−46	111	58	−47	56	0	−49	0	−49	0	0	0	−49	1	−49	−49	0
Tyr	0	−22	128	−24	−24	0	−25	0	−75*	−75*	0	0	68	0	−27	20	−51	−51	−27	−27

A DP value of 0 indicates that the particular amino acid is not present at the position in the selected data set.
*Statistically significant DP values.

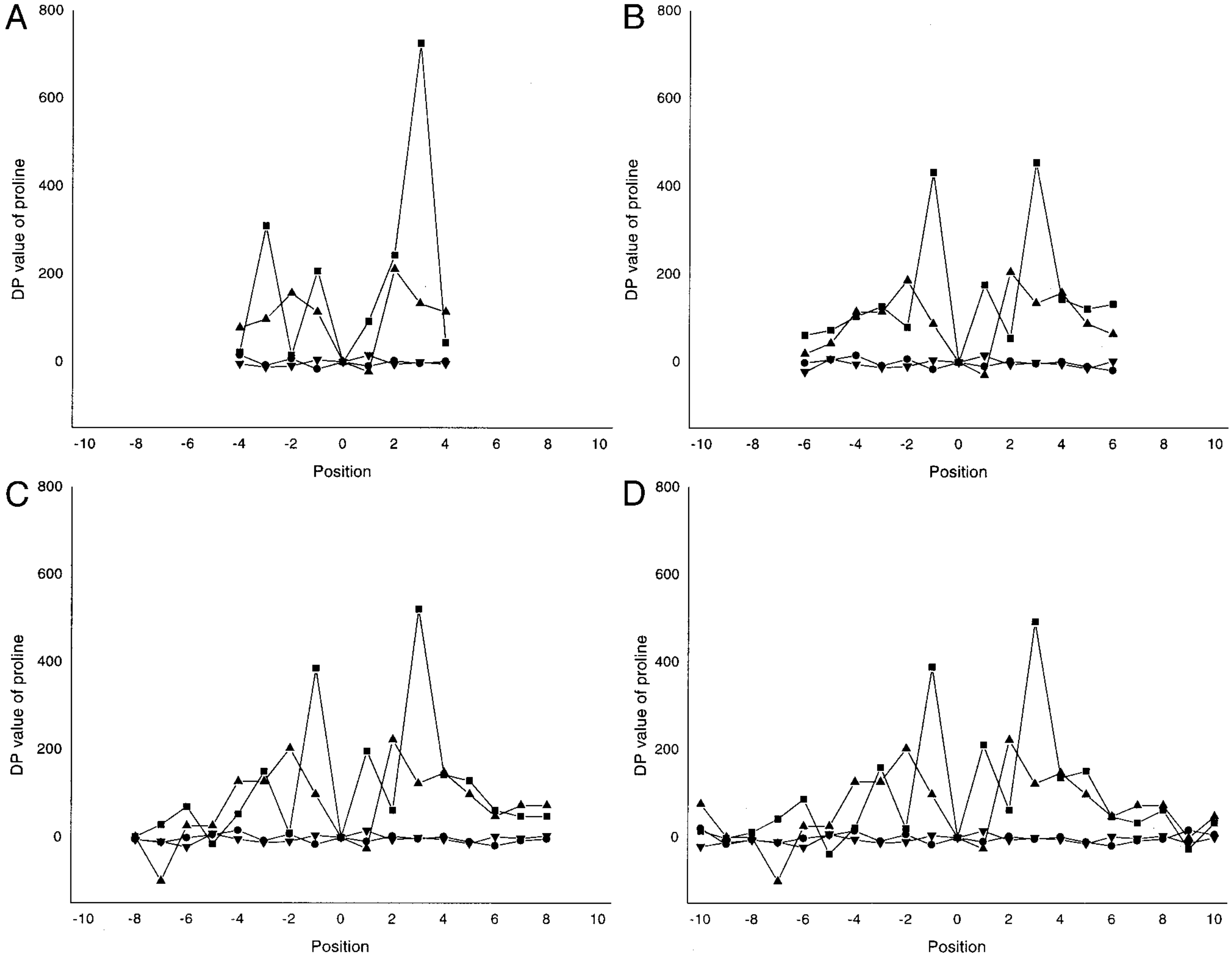

FIGURE 2 Positional preference of proline around the *O*-glycosylation site with different window sizes. ■, DP value of Pro around glycosylated Thr; ●, DP value of Pro around nonglycosylated Thr; ▲, DP value of Pro around glycosylated Ser; ▼, DP value of Pro around nonglycosylated Ser. The window size is varied from 7 to 21. Plots are for the preference of proline at positions from −4 to +4 (*A*), −6 to +6 (*B*), −8 to +8 (*C*), and −10 to +10 (*D*). DP values of proline with respect to Ser and Thr glycosylation are calculated separately.

the sequences contain Thr-Ala-Pro-Pro in which Thr is the glycosylated one but Ser-Ala-Pro-Pro is not found. Also a search was carried out in more than 500 nonhomologous proteins, excluding *O*-glycosylated glycoproteins, from Brookhaven Protein Data Bank for the sequence motif Thr-Ala-Pro-Pro. It was found that no such sequence motif is present in these nonglycosylated proteins. It was also found that this sequence motif is part of a tandem repeat in a single protein, human MUC1. Also among the proline-containing (at +3) sequence motifs, Thr-Val-X-Pro, Ser/Thr-Pro-X-Pro, and Thr-Ser-Ala-Pro are preferred favorably (8%, 15%, and 15%), where X can be any amino acid.

Mucin-type glycosylation

As discussed earlier, in 97 proteins (54%) mucin-type glycosylation occurs, in which the carbohydrate moiety attached to the hydroxyamino acids is GalNAc. There are 181 Ser glycosylation sites and 463 Thr glycosylation sites (65% of the total glycosylation sites). The multiple and isolated glycosylation sites were separated out, and it was found that the number of isolated (single glycosylation) sites (76 glycosylation sites) is much fewer than that of the multiple glycosylation sites (588 glycosylation sites). This indicates that multiple glycosylation is common in this type. The deviation parameter calculations were carried out for multiple glycosylation sequences (Table 3) as the single glycosylation sequences lack sufficient data for statistical interpretation.

Around multiple glycosylation sites the amino acids preferred at various positions are Pro, Ala, Ser, Thr, Asp, Glu, Gly, His, Arg, and Val. It is seen that proline is preferred at various positions (−9, −7, −6, −5, −3, −1, +2, +3, +6, +7, and +9). The DP value of Pro is maximal at the +3 position (289) followed by its value at −5 (186), +7 (172),

TABLE 3 Deviation parameters for the amino acids at positions from −10 to +10 for the multiple glycosylation sequences in mucin-type glycosylation

Amino	DP values at positions																			
acid	*i* − 10	*i* − 9	*i* − 8	*i* − 7	*i* − 6	*i* − 5	*i* − 4	*i* − 3	*i* − 2	*i* − 1	*i* + 1	*i* + 2	*i* + 3	*i* + 4	*i* + 5	*i* + 6	*i* + 7	*i* + 8	*i* + 9	*i* + 10
Ala	254*	58*	−8	95*	82*	−2	224*	−10	151*	158*	92*	197*	144*	155*	2	−34*	62*	72*	−25	67*
Cys	−71*	−88*	−71*	−77*	−88*	−88*	−94*	−94*	−94*	0	−94*	−94*	−83*	−94*	−89*	−78*	−78*	−78*	−62*	−83*
Asp	−60*	−11	195*	47	−40*	−65*	82*	168*	33	−93*	−62*	−45*	−83*	79*	72*	55	−62*	−45*	69*	208*
Glu	−33	−40*	−33	−47*	−31	7	−31	90*	−16	−64*	97*	71*	−77*	−20	−1	−4	−30	−1	55*	74*
Phe	−82*	−58*	−35	−94*	−59*	−88*	−65*	−71*	−65*	−54*	−83*	−66*	−54*	−66*	−88*	−49*	−66*	−71*	−32	−60*
Gly	47*	23	34	68*	−42*	32	27	−54*	158*	−53*	42	−35*	−59*	49*	29	124*	−24	46*	46*	40
His	−37	−51*	−37	−38	−72*	−59*	−45	250*	−53*	−60*	−60*	−27	−66*	−46*	251*	0	−7	−46*	−60*	−40
Ile	−49*	−9	−35	−60*	−20	−21	−36	−41	20	−18	−32	−52*	−52*	−18	−56*	−42*	−52*	−52*	−56*	−23
Lys	−40	−77*	−54*	−46*	−50*	−42*	−46*	−55*	−60*	−78*	−69*	−65*	−91*	−52*	−56*	−52*	−43*	−39	0	−69*
Leu	−50*	−53*	−48*	−81*	−72*	−58*	−49*	−59*	−63*	−82*	−68*	−62*	−66*	−50*	−44*	−55*	−68*	−62*	−68*	−70*
Met	−75*	−39	−64*	−28	−5	−41	5	−30	−7	−53	−42	−31	−19	−65*	−31	−77*	−42	3	−42	−65*
Asn	−68*	−78*	−52*	−42	−68*	−63*	−58*	−53*	−48*	−64*	−79*	−49*	−54*	−59*	−59*	−74*	−54*	−59*	−44*	−69*
Pro	−10	92*	−7	72*	150*	186*	6	98*	25	72*	28	144*	289*	24	8	73*	172*	−11	81*	12
Gln	−38	−43*	−65*	−26	−44*	−57*	−66*	−23	−57*	−62*	−37	−62*	−54*	−50*	−25	−54*	−54*	−50*	−66*	−54*
Arg	−72*	−68*	−76*	−68*	90*	−61*	−77*	−84*	−84*	−62*	−66*	−43*	−92*	−66*	−51*	110*	−69*	−73*	−58*	−36
Ser	95*	70*	2	2	64*	41*	52*	7	1	92*	135*	17	61*	−16	−20	−7	64*	47*	131*	32
Thr	11	91*	173*	123*	23	94*	5	1	−18	83*	27	−7	28	80*	156*	28	9	152*	17	3
Val	−22	−17	−21	−46*	−26	−42*	−30	−12	−13	96*	−26	−36*	−46*	−28	−38*	−33*	69*	−6	−46*	−41*
Trp	−85*	−56	−85*	−85*	−56	−71*	−85*	−85*	−85*	−86*	−72*	−72*	−72*	−72*	0	−72*	0	−58	−58	−86*
Tyr	−80*	−80*	−73*	−93*	−87*	−67*	−87*	−61*	−68*	−81*	−81*	−62*	−49*	−74*	−81*	−68*	−62*	−24	−49*	−43

A DP value of 0 indicates that the particular amino acid is not present at the position in the selected data set.
*Statistically significant DP values.

−6 (150), and +2 (144) positions. It is noted that Pro is less preferred at +1 and +5 positions when compared with the computations carried out using the whole database. Ala and Gly possess positive significant DP values at positions −10, −7, −2, +4, and +8 in common (Table 3) and Val is preferred at −1 and +7 positions. It is noted that amino acids with small side chains are preferred around multiple glycosylation sites. The hydroxyamino acids Ser and Thr are also favored around multiple glycosylation sites at various positions (Table 3). It is interesting to note that the acidic amino acids Asp and Glu are preferred at various positions around the glycosylated Ser/Thr, a trend that is not found when the glycosylation sites from the whole database are considered. This clearly indicates that acidic amino acids may have a structural role in mucin-type glycosylation. In addition, His is preferred at −3 and +5 positions and Arg at positions −6 and +6 (Table 3).

O-glycosylation relative to *N*-glycosylation

In *N*-linked glycosylation, it is known that Asn-X-Ser/Thr is the consensus sequence (Hunt and Dayhoff, 1970; Marshall, 1972), and it is an essential but not a sufficient condition for *N*-glycosylation to occur. A proline residue at X position prevents *N*-glycosylation and also a proline beyond Ser/Thr (+3 position) inhibits *N*-glycosylation (Gavel and von Heijne, 1990), whereas our calculations show that in *O*-glycosylation, proline residues are present close to the *O*-glycosylated Ser/Thr and it enhances *O*-glycosylation whenever it is present nearer to Ser/Thr especially at −1 and +3 positions. It is also known that in *N*-glycosylation adjacent and consecutive residues are glycosylated in low numbers. It may be that steric hindrance prevents more closely spaced sites from being glycosylated at the same time (Gavel and von Heijne, 1990). In *O*-glycosylation, consecutive and adjacent serine and threonine residues are glycosylated, and clusters of *O*-glycosylation sites were noted in 98 glyco-proteins (of 180). It is also known that Asn-X-Ser and Asn-X-Thr sequences in proteins are doubly glycosylated. The Asn and the Ser or Thr in the tripeptide are both glycosylated, for example, in human tissue kallikrein (Kellermann et al., 1989), human coagulation factor (Iwanaga et al., 1990), glycoprotein in Friend murine leukemia virus (Chen, 1982), and pig glycophorin (Tomita and Marchesi, 1975). Aromatic amino acids are less preferred near the site of *O*-glycosylation (Tables 1–3). This also contrasts with what is observed in *N*-glycosylation, where the aromatic amino acids are preferred at X position and close to the glycosylation site, and are believed to stabilize the structure due to the stacking interactions of the aromatic ring with the glycan (Christlet et al., 1999; Imberty and Perez, 1995).

CONCLUSION

The data presented in the present work clearly indicate that there is a pronounced positional preference for the amino acids at various positions around the *O*-glycosylation site.

Pro occurs preferentially at many positions close to the site of glycosylation and, in particular, strongly favors *O*-glycosylation when it is in the −1 and/or +3 positions. This may indicate that proline plays a structural role in directing *O*-glycosylation in contrast to the negative role it plays in *N*-glycosylation. Proline more frequently occurs at the −2 and +2 positions when the site of glycosylation is a Ser residue. In addition, Ser and Thr are preferred around the multiple glycosylation sites probably due to the effect of clusters of closely spaced *O*-glycosylation sites. Around multiple glycosylation sites, the other amino acids preferred favorably are Ala, Gly, Asp, His, and Val. Cysteine, in positions close to the site of glycosylation, hampers *O*-glycosylation. The aromatic amino acids and amino acids with bulky side chains also hinder *O*-glycosylation. Some potential sequence motifs such as Thr-Ala-Pro-Pro, Thr-Val-X-Pro, Ser/Thr-Pro-X-Pro, and Thr-Ser-Ala-Pro occur frequently in the data set. In mucin-type glycosylation, in which the sugar attached to the hydroxyamino acids is GalNAc, the same calculations show likely similar results and also Asp and Glu residues are highly preferred, indicating a structural role for these acidic amino acids. In future it will be of much interest to investigate further the possible structural and conformational implications of some of these suggested positional preferences of the various amino acids around the site of glycosylation. This is a potentially important study, and such analyses will surely contribute an important part of our knowledge base in the future on *O*-glycosylation sites. These results will be of interest to molecular biologists and protein engineers to identify *O*-glycosylation sites important in molecular recognition processes.

We thank Dr. Margaret Biswas for many helpful discussions and assistance with this work. The financial assistance received from the Department of Biotechnology for this project is gratefully acknowledged. T.H.T.C. thanks the University Grants Commission for JRF.

REFERENCES

Allen, A. K., N. N. Desai, A. Neuberger, and J. M. Creeth. 1978. Properties of potato lectin and the nature of its glycoprotein linkages. *Biochem. J.* 171:665–674.

Aubert, J.-P., G. Biserte, and M.-H. Loucheux-Lefebvre. 1976. Carbohydrate-peptide linkage in glycoproteins. *Arch. Biochem. Biophys.* 175: 410–418.

Bause, E. 1983. Structural requirements of *N*-glycosylation of proteins. *Biochem. J.* 209:331–336.

Bause, E., and H. Hettkamp. 1979. Primary structural requirements for *N*-glycosylation of peptides in rat liver. *FEBS Lett.* 108:341–344.

Carraway, K., and S. Hull. 1991. Cell surface mucin-type glycoproteins and mucin-like domains. *Glycobiology.* 1:131–138.

Chen, R. 1982. Complete amino acid sequence and glycosylation sites of glycoprotein gp71A of Friend murine leukemia virus. *Proc. Natl. Acad. Sci. U.S.A.* 79:5788–5792.

Chou, K.-C. 1995. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. *Protein Sci.* 4:1365–1383.

Christlet, T. H. T., M. Biswas, and K. Veluraja. 1999. A database analysis of the potential glycosylating Asn-X-Ser/Thr consensus sequences. *Acta Cryst.* D55:1414–1420

Dahms, N. M., and G. W. Hart. 1986. Influence of quaternary structure on *O*-glycosylation. *J. Biol. Chem.* 261:13186–13196.

Elhammer, A. P., R. A. Poorman, E. Brown, L. L. Maggiora, J. G. Hoogerheide, and F. J. Kezdy. 1993. The specificity of UDP-GalNAc: polypeptide *N*-acetylgalactosaminyltransferase as inferred from a database of in vivo substrates and from the in vitro glycosylation of proteins and peptides. *J. Biol. Chem.* 268:10029–10038.

Elliott, S., T. Bartley, E. Delorme, P. Derby, R. Hunt, T. Lorenzini, V. Parker, M. F. Rohde, and K. Stoney. 1994. Structural requirements for addition of *O*-linked carbohydrate to recombinant erythropoietin. *Biochemistry.* 33:11237–11245.

Fiat, A.-M., J. Jolles, J. P. Aubert, M.-H. Loucheux-Lefebvre, and P. Jolles. 1980. Localisation and importance of the sugar part of human casein. *Eur. J. Biochem.* 111:333–339.

Fukuda, M. 1991. Leukosialin, a major *O*-glycan containing sialoglycoprotein defining leukocyte differentiation and malignancy. *Glycobiology.* 1:347–356.

Gavel, Y., and G. von Heijne. 1990. Sequence differences between glycosylated and nonglycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng.* 3:433–442.

Gerken, T. A., C. L. Owens, and M. Pasumarthy. 1997. Determination of site specific *O*-glycosylation pattern of the porcine submaxillary mucin tandem repeat. *J. Biol. Chem.* 272:9709–9719.

Gooley, A. A., B. J. Classon, R. Marschalek, and K. L. Williams. 1991. Glycosylation sites identified by detection of glycosylated amino acids released from Edman degradation: the identification of Xaa-Pro-Xaa-Xaa as a motif for threonine *O*-glycosylation. *Biochem. Biophys. Res. Commun.* 178:1194–1201.

Gupta, R., H. Birch, K. Rapacki, S. Brunak, and J. E. Hansen. 1999. O-GLYCBASE version 4.0: a revised database of *O*-glycosylated proteins. *Nucleic Acids Res.* 27:370–372.

Haltiwanger, R. S., W. G. Kelly, E. P. Roquemore, M. A. Blomberg, L. Y. Dong, L. Kreppel, T. Y. Chou, and G. W. Hart. 1992. Glycosylation of nuclear and cytoplasmic proteins is ubiquitous and dynamic. *Biochem. Soc. Trans.* 20:264–269.

Hansen, J. E., O. Lund, J. Engelbrecht, H. Bohr, J. O. Nielsen, J.-E. S. Hansen, and S. Brunak. 1995. Prediction of *O*-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc: polypeptide *N*-acetylgalactosaminyltransferase. *Biochem. J.* 308:801–813.

Hansen, J. E., O. Lund, J. Nilsson, K. Rapacki, and S. Brunak. 1998. O-GLYCBASE version 3.0: a revised database of *O*-glycosylated proteins. *Nucleic Acids Res.* 26:387–389.

Hansen, J. E., O. Lund, K. Rapacki, and S. Brunak. 1997. O-GLYCBASE version 2.0: a revised database of *O*-glycosylated proteins. *Nucleic Acids Res.* 25:278–282.

Hart, G. 1992. Glycosylation. *Curr. Opin. Cell. Biol.* 4:1017–1023.

Hausler, A., L. Ballou, C. E. Ballou, and P. W. Robbins. 1992. Yeast glycoprotein biosynthesis: MNT1 encodes an alpha-1,2-mannosyltransferase involved in *O*-glycosylation. *Proc. Natl. Acad. Sci. U.S.A.* 89: 6846–6850.

Hunt, L. T., and M. O. Dayhoff. 1970. The occurrence of proteins of the tripeptides Asn-X-Ser and Asn-X-Thr and of bound carbohydrate. *Biochem. Biophys. Res. Commun.* 39:757–765.

Imberty, A., and S. Perez. 1995. Stereochemistry of the *N*-glycosylation sites in glycoproteins. *Protein Eng.* 8:699–709.

Iwanaga, S., H. Nishimura, S. Kawabata, W. Kisiel, S. Hase, and T. Ikenaka. 1990. A new trisaccharide sugar chain linked to a serine residue in the first EGF-like domain of clotting factors VII and IX and protein Z. *Adv. Exp. Med. Biol.* 281:121–131.

Jentoft, N. 1990. Why are proteins glycosylated? *Trends Biochem. Sci.* 15:291–294.

Kellermann, J., F. Lottspeich, R. Geiger, and R. Deutzmann. 1989. Human urinary kallikrein: amino acid sequence analysis and carbohydrate attachment. *Adv. Exp. Med. Biol.* 247A:519–525.

Lehle, L., and E. Bause. 1984. Primary structural requirements for *N*- and *O*- glycosylation of yeast mannoproteins. *Biochim. Biophys. Acta.* 799: 246–251.

Marshall, R. D. 1972. Glycoproteins. *Annu. Rev. Biochem.* 41:673–702.

Muller, S., S. Goletz, N. Packer, A. Gooley, A. M. Lawson, and F.-G. Hanisch. 1997. Localization of *O*-glycosylation sites on glycopeptide fragments from lactation associated MUC1. *J. Biol. Chem.* 272: 24780–24793.

Nishimura, H., T. Takao, S. Hase, Y. Shimonishi, and S. Iwanaga. 1992. Human factor IX has a tetrasaccharide *O*-glycosidically linked to serine 61 through the fucose residue. *J. Biol. Chem.* 267:17520–17525.

O'Connell, B., F. K. Hagen, and L. A. Tabak. 1992. The influence of flanking sequences on the *O*-glycosylation of threonine in vitro. *J. Biol. Chem.* 267:25010–25018.

O'Connell, B., L. A. Tabak, and N. Ramasubbu. 1991. The influence of flanking sequences on *O*-glycosylation. *Biochem. Biophy. Res. Commun.* 180:1024–1030.

Pisano, A., J. Redmond, K. Williams, and A. Gooley. 1993. Glycosylation sites identified by solid-phase Edman degradation: *O*-linked glycosylation motifs on human glycophorin A. *Glycobiology.* 3:429–435.

Spiro, R. 1973. Glycoproteins. *Adv. Protein Chem.* 27:349–467.

Strous, G., and J. Dekker. 1992. Mucin-type glycoproteins. *Crit. Rev. Biochem. Mol. Biol.* 27:57–92.

Tomita, M., and V. T. Marchesi. 1975. Amino acid sequence and oligosaccharide attachment sites of human erythrocyte glycophorin. *Proc. Natl. Acad. Sci. U.S.A.* 72:2964–2968.

Wang, Y., N. Agrwal, A. E. Eckhardt, R. D. Stevens, and R. L. Hills. 1993. The acceptor substrate specificity of porcine submaxillary UDP-GalNAc: polypeptide *N*-acetylgalactosaminyltransferase is dependent on the amino acid sequences adjacent to serine and threonine residues. *J. Biol. Chem.* 268:22979–22983.

Wilson, I. B. H., Y. Gavel, and G. von Heijne. 1991. Amino acid distributions around *O*-linked glycosylation sites. *Biochem. J.* 275:529–534.

Yanagishita, M., and V. C. Hascall. 1992. Cell surface heparan sulfate proteoglycans. *J. Biol. Chem.* 267:9451–9454.

Yoshida, A., M. Suzuki, H. Ikenaga, and M. Takeuchi. 1997. Discovery of shortest sequence motif for high level mucin type *O*-glycosylation. *J. Biol. Chem.* 272:16884–16888.

Young, J. D., D. Tsuchiya, D. E. Sandlin, and M. J. Holroyde. 1979. Enzymatic *O*-glycosylation of synthetic peptides from sequences in basic myelin protein. *Biochemistry.* 18:4444–4448.