Glycobiology vol. 14 no. 2 pp. 103–114, 2004
DOI: 10.1093/glycob/cwh008
Advance Access publication September 26, 2003

Statistical analysis of the protein environment of *N*-glycosylation sites: implications for occupancy, structure, and folding

Andrei-J. Petrescu[2,3], Adina-L. Milac[3], Stefana M. Petrescu[2,3], Raymond A. Dwek[2], and Mark R. Wormald[1,2]

[2]Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, U.K.; and [3]Institute of Biochemistry of the Romanian Academy, 296 Spl. Independentei 296, 77700 Bucharest 17, Romania

Received on July 17, 2003; revised on September 9, 2003; accepted on September 10, 2003

We recently reported statistical analysis of structural data on glycosidic linkages. Here we extend this analysis to the glycan–protein linkage, and the peptide primary, secondary, and tertiary structures around *N*-glycosylation sites. We surveyed 506 glycoproteins in the Protein Data Bank crystallographic database, giving 2592 glycosylation sequons (1683 occupied) and generated a database of 626 nonredundant sequons with 386 occupied. Deviations in the expected amino acid composition were seen around occupied asparagines, particularly an increased occurrence of aromatic residues before the asparagine and threonine at position +2. Glycosylation alters the asparagine side chain torsion angle distribution and reduces its flexibility. There is an elevated probability of finding glycosylation sites in which secondary structure changes. An 11-class taxonomy was developed to describe protein surface geometry around glycosylation sites. Thirty-three percent of the occupied sites are on exposed convex surfaces, 10% in deep recesses and 20% on the edge of grooves with the glycan filling the cleft. A surprisingly large number of glycosylated asparagine residues have a low accessibility. The incidence of aromatic amino acids brought into close contact with the glycan by the folding process is higher than their normal levels on the surface or in the protein core. These data have significant implications for control of sequon occupancy and evolutionary selection of glycosylation sites and are discussed in relation to mechanisms of protein fold stabilization and regional quality control of protein folding. Hydrophobic protein–glycan interactions and the low accessibility of glycosylation sites in folded proteins are common features and may be critical in mediating these functions.

Key words: glycan–protein linkage/*N*-glycosylation sites/occupancy/protein folding/X-ray diffraction

Introduction

Carbohydrates attached to proteins can play a wide variety of roles. A major function of protein-linked glycans is to provide additional recognition epitopes for protein receptors (Drickamer and Taylor, 1993; Lis and Sharon, 1998). Such recognition events are involved in a wide range of processes, including protein trafficking, initiation of inflammation, and host defense (Lasky, 1995; Maly *et al*., 1996; Crocker, 2002). In these cases, the location of the glycan attachment site on the protein is unlikely to be critical to function but will depend on the precise sequence of the glycan (Gagneux and Varki, 1999). Of particular interest is the role of glycan recognition in the glycoprotein chaperone-assisted folding and quality control mechanisms (Parodi, 2000a,b). There is increasing evidence that this process also depends on the precise location of the *N*-glycosylation sites because the quality control mechanism appears to be regional and so not all glycans are equally important in the folding process (Petrescu *et al*., 2000; Daniels *et al*., 2003).

Considerable work has been done to characterize the sequences of oligosaccharides attached to proteins (Rudd and Dwek, 1997; Duus *et al*., 2000) and to determine their 3D structures (Imberty and Perez, 2000). Databases are available for glycan primary structures (Cooper *et al*., 2001), and we have recently reported on the statistical analysis of the available X-ray diffraction data on oligosaccharide linkage conformations (Petrescu *et al*., 1999; Wormald *et al*., 2002).

In addition to their role as recognition markers, the size and hydrophilicity of glycans or simply their presence at definite locations can alter the behavior of glycoproteins, making them more soluble or stable, protecting them locally from proteolysis or aggregation, or masking their antigenic sites (Dwek, 1996; Rudd *et al*., 1999; Wormald and Dwek 1999; Helenius and Aebi, 2001). However, there have been few systematic studies of the locations of such sites in proteins. There is a database of *O*-glycosylated proteins (Gupta *et al*., 1999) and statistical analysis has been performed on the sequences around such sites to identify preferential motifs for *O*-glycosylation (Christlet and Veluraja, 2001). Potential *N*-glycosylation sites can be identified by the presence of the Asn-X-Ser/Thr sequon in peptide sequence databases (Apweiler *et al*., 1999). However, it is very difficult to determine which of these sites are occupied, even if the protein in question is known to be glycosylated. Structural analysis by X-ray crystallography provides direct and unambiguous evidence for the occupancy of a glycosylation site. Evidence from X-ray crystallography for the unoccupancy of a site is more ambiguous because the absence of the glycan is only one reason for the absence of resolved electron density.

Some results on the analysis of crystallographic *N*-glycosylation sites are available. Imberty and Perez (1995) surveyed the stereochemistry of 44 *N*-glycosylation sites.

[1]To whom correspondence should be addressed; e-mail: mark@glycob.ox.ac.uk

They concluded that modified Asn residues showed the same distribution of conformations as observed for unmodified residues, and that 25% of *N*-glycosylation sites occurred on β-turns. Veluraja and co-workers surveyed 696 *N*-glycosylation sites (Christlet *et al.*, 1999). They concluded that the residues Gly, Asn, and Phe occur more frequently at position X and analyzed in detail the backbone conformations of the Asn-X-Ser/Thr sequons.

In this article, we survey 2592 glycosylation sequons in glycoproteins, defined as proteins in which at least one occupied glycosylation site is observed directly in the crystal structure, 1683 of which are occupied. We analyze the conformational preferences of the Asn side chain and Asn-glycan linkage on this full data set. A subset of 626 nonredundant sequons, 386 of which are occupied, was also generated. On this subset, we analyze the statistical distribution of amino acids in a 35-residue stretch around the glycosylation site, the secondary structure in which the *N*-glycosylation sites occur, and the protein surface topology and properties around the glycosylation site.

The statistical analysis of the sequence around glycosylation sites can provide insights into the selectivity of the oligosaccharyl transferase (OST) complex. Because glycosylation is cotranslational, the secondary and tertiary structure around a glycosylation site will have no direct effect on whether that site is occupied. However, the statistical analZysis of the secondary and tertiary structure can provide insights into the types of *N*-glycosylation sites that have been selected and retained during evolution.

Results

A full data set of glycoprotein structures in which at least one *N*-linked glycan is resolved was obtained by exhaustive search of the Brookhaven Protein Data Bank (PDB) (Berman *et al.*, 2000). The complete set of putative *N*-glycosylation sites was obtained by scanning this set of glycoproteins for Asn-X-Ser/Thr sequences. These were classified as occupied based on the presence of a GlcNAc residue at that site in the crystal structure. Many files examined in this survey contained identical or very closely related glycoproteins, crystallized in various conditions by different laboratories, or multiple copies of a glycoprotein in the unit cell. To obtain a nonredundant set the complete set of *N*-glycosylation sites was clustered in families followed by picking one representative member from each family as discussed in *Materials and methods*. The full data set contained 2592 sequons from 506 structures, and the nonredundant set contained 626 sequons from 216 structures (Table I). A site identified as unoccupied in these structures may be the result of the initial absence of a glycan (true unoccupancy), sample processing prior to crystallization, or the absence of interpretable electron density and so have to be treated as a set enriched in unoccupied sites rather than a pure set of unoccupied sites. Therefore we focus only on the occupied sites, 1683 in total, of which 386 nonredundant. Conformational analysis was performed on the full set (each structure being an independent experimental data set), whereas amino acid sequence, secondary structure, and surface property analyses were performed on the nonredundant set to avoid bias toward overrepresented protein families.

Table I. Public domain glycoprotein structures containing *N*-linked oligosaccharides

	Full	Nonredundant
No. of crystal structures with *N*-linked glycans	506	216
No. of peptide chains	929	253
No. of sequons	2592	626
No. of occupied sequons	1683	386
% Occupied	64.9	61.7

The full data set consists of all structures with at least one GlcNAc residue linked to an Asn residue. The nonredundant data set consists of a subset of the full data set containing one representative from each occupied glycosylation site structural family (see *Materials and methods* for details).

Occupation of N-*glycosylation sites*

Just under 65% of the potential glycosylation sequons are occupied in the full data set and just over 60% in the nonredundant data set (Table I). Of these, approximately 70% are of the type Asn-X-Thr and 30% of the type Asn-X-Ser. This ratio is reversed in the set of unoccupied glycosylation sites with 41% Asn-X-Thr and 59% Asn-X-Ser. No occupied sequons of the type Asn-Pro-Ser/Thr were observed.

Amino acid sequence analysis around occupied N-*glycosylation sites*

The probability of finding each amino acid at each sequence position around occupied and unoccupied glycosylation sites, from −9 to +25, has been calculated for the structures in the nonredundant data set. This region was chosen to include residues immediately around the glycosylation site (±9) and downstream residues that may interact with the translocon complex while glycosylation occurs (glycosylation occurring approximately 30 residues from the ribosome; Varki *et al.*, 1999). The probabilities of finding proline, a hydrophobic amino acid, a hydrophilic amino acid, a basic amino acid or an acidic amino acid at each sequence position are shown in Figure 1A and compared to the probabilities of finding such amino acids at any point in the peptide sequence. As can be seen, the only significant differences between these two probabilities occur in the immediate vicinity of the glycosylation site (positions −6 to +4).

There is a marked preference for hydrophobic amino acids either side of the glycosylation site (Figure 1A). This preference is not uniform among hydrophobic amino acids. There is an increase in the probability of finding aromatic residues immediately before a glycosylation site, whereas there is an increased probability of finding small hydrophobic amino acids at position +1 and larger hydrophobic amino acids at position +3 (Figure 1B). It is interesting to note that there are corresponding decreases in the probabilities of finding an aromatic or large aliphatic hydrophobic

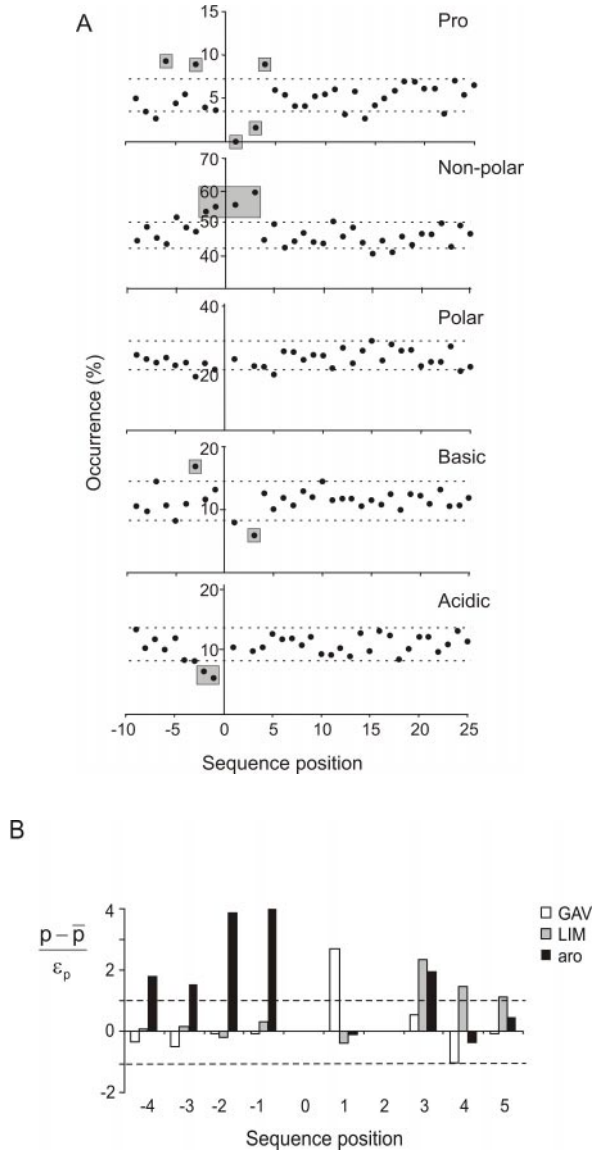

Fig. 1. (**A**) Amino acid type distribution around occupied glycosylation sites. The percentage occurrence is plotted versus sequence position for positions −10 to +25 around an occupied glycosylation site. Residues specified by the glycosylation sequon (i.e., position 0: Asn, and position +2: Ser or Thr) are not plotted. The horizontal lines show plus and minus one standard deviation for the average percentage occurrence (see *Materials and methods* for details). Pro, proline; Nonpolar, hydrophobic amino acids (Gly, Ala, Val, Leu, Ile, Met, Phe, Tyr, Trp); Polar, hydrophilic amino acids (Ser, Thr, Asn, Gln); Basic, basic amino acids (Lys, Arg, His); Acidic, acidic amino acids (Asp, Glu). (**B**) Variation of the composition of light nonpolar (GAV = Gly, Ala, and Val), bulky hydrophobic (LIM = Leu, Ile, and Met), and aromatic amino acids (aro = Phe, Tyr, and Trp) in the region −4 to +5 around an occupied glycosylation site, measured in standard deviations relative to the average value. *p*, percentage occurrence around an occupied glycosylation site; $\bar{p}$ average percentage occurrence; ε_p, standard deviation for the average percentage occurrence.

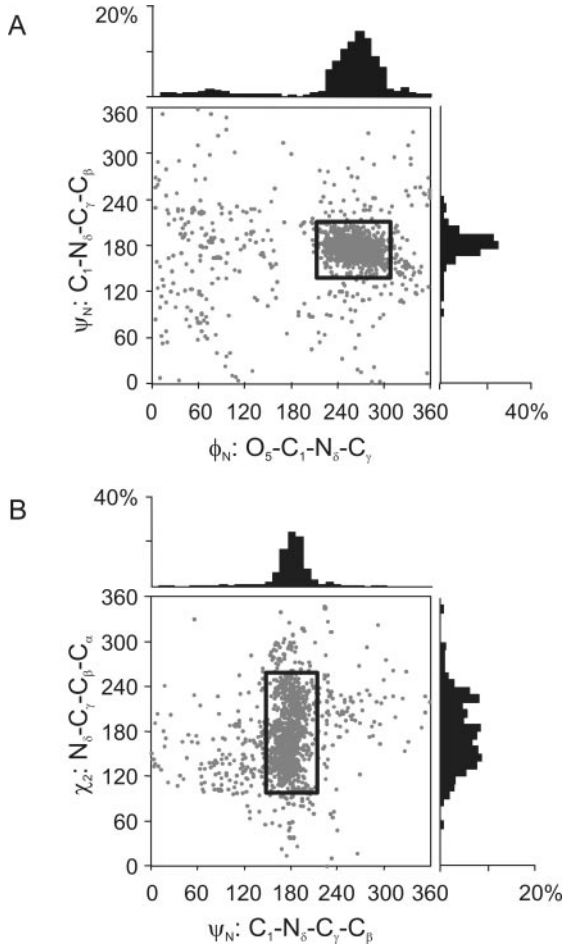

Fig. 2. Torsion angle and histogram plots for the Asn-GlcNAc linkage. Structures associated with a distinct conformer are shown by the boxes. (**A**) Plot of ϕ_N versus ψ_N. (**B**) Plot of ψ_N versus χ_2.

residue preceding unoccupied glycosylation sites (data not shown).

There is an increase in the probability of finding proline in the vicinity of a glycosylation site except for the complete absence of proline at position +1 (already noted) and a reduced probability at position +3. There are no significant alterations in the probabilities of finding a hydrophilic residue near an occupied glycosylation site and few alterations for basic residues. However, there is a notable reduction in the probability of finding acidic residues immediately before the glycosylation site. Again this is accompanied by an increased probability of finding acidic residues immediately before unoccupied glycosylation sites.

Conformation of the protein–glycan linkage

The torsion angles, ϕ_N and ψ_N, for the Asn-GlcNAc linkage (Figure 2) were measured for all the structures in the full data set of occupied glycosylation sites, as in this case each individual entry in the PDB is a separate experimental result. The side chain torsion angles, χ_1 and χ_2, were measured for four subsets of Asn residues from the full data set: (1) Asn residues in occupied sequons (Figure 3A), (2) Asn residues in unoccupied sequons, (3) Asn residues not in glycosylation sequons and with a solvent accessibility greater than zero (i.e., surface residues, Figure 3B), and (4) Asn residues not in glycosylation sequons and with a

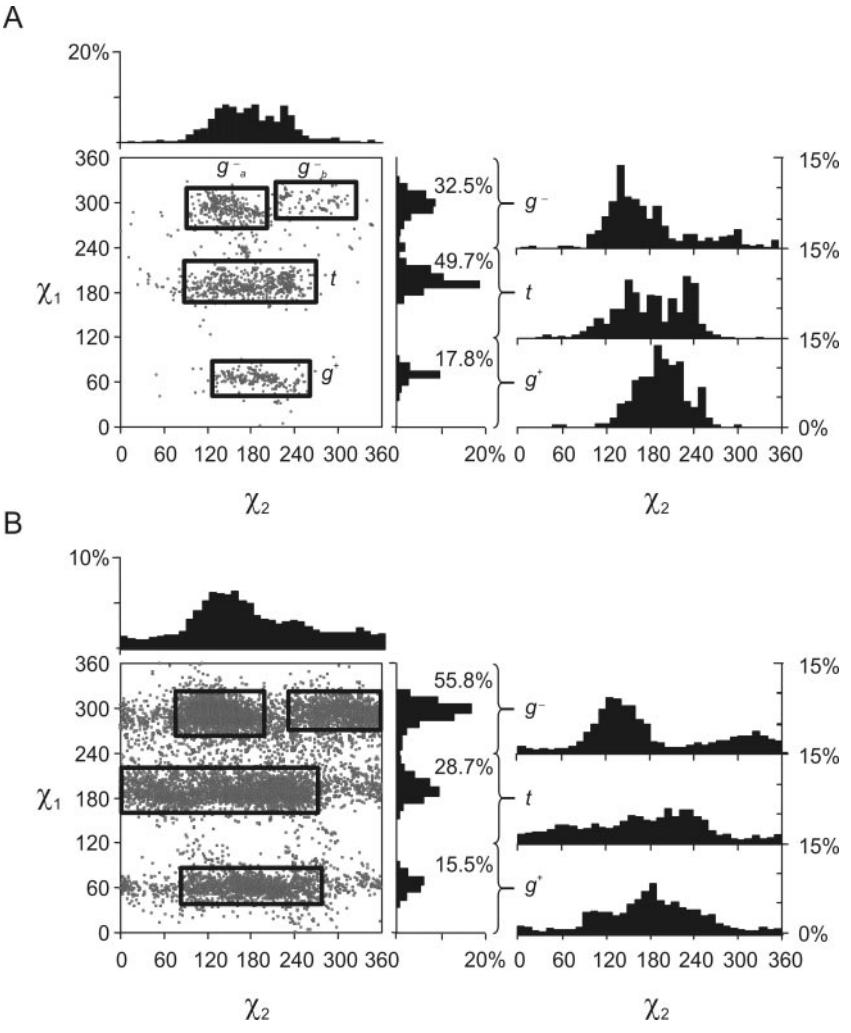

Fig. 3. Torsion angle and histogram plots of χ_1 versus χ_2 for Asn residue side chain subsets. The χ_2 histogram plot is shown for each distinct χ_1 conformer, as well as for the total population. (**A**) Asn residues in occupied glycosylation sequons. (**B**) Asn residues not in glycosylation sequons and with a solvent accessibility greater than zero.

solvent accessibility of zero (i.e., buried residues). The statistical results are summarized in Table II.

As can be seen from Figure 2A, the Asn-GlcNAc linkage only adopts one significantly populated conformation at $\phi_N/\psi_N = 260.2° \pm 21.1°/176.8° \pm 12.0°$. The ψ_N distribution is much narrower than the ϕ_N distribution. Only 21 of the 1678 structures showed a cis amide bond (ψ_N between $-30°$ and $+30°$).

The side chains of the glycosylated Asn residues fall in to three well-defined conformations (Figure 3A) with χ_1 values of 60°, 180°, and 300° and χ_2 of 180°, the latter showing a wide distribution. The three subsets of unmodified Asn residues all have very similar conformational behavior and show the same three conformers observed for the glycosylated subset. However, there are two very noticeable differences between the glycosylated subset and the other three. First, the width of the χ_2 distribution is much smaller for the glycosylated Asn residues, the standard deviation reducing on average from 82.1° to 52.5° for all three conformers on glycosylation. Second, the relative populations of the three conformers are different for the glycosylated subset. The subsets of unmodified side chains all show a preference for the g^- conformer ($\chi_1 = 300°$, 55.8%) over the t conformer ($\chi_1 = 180°$, 28.7%) with the g^+ conformer ($\chi_1 = 60°$) only present at 15.5%. For the glycosylated subset, the g^+ conformer is still small, 18.3%, but the t conformer is now preferred (50.1%) compared to g^- (31.6%).

Secondary structure around occupied N*-glycosylation sites*

Figure 4A shows the analysis of secondary structure around glycosylated Asn residues in the nonredundant data set. As can be seen, occupied glycosylation sites can occur on all forms of secondary structure. There is a bias in favor of turns and bends: ~27% and 17.5% of occupied sites are on turns and bends, respectively, whereas the overall incidence of these secondary structures in the nonredundant data set are 14% and 11%, respectively. The lowest incidence of

Table II. Average torsion angles and standard deviations for all distinct conformers of the Asn-GlcNAc linkage and Asn side chain from the full data set

	No. structures in distinct conformers	Conformer	(%)	χ_1	χ_2	ψ_N	ϕ_N
Asn in occupied sequon	1292	g^+	17.8	63.6 ± 8.9	191.1 ± 31.6	178.5 ± 13.9	253.7 ± 21.5
		t	49.7	191.6 ± 14.4	177.6 ± 43.0	177.3 ± 12.3	261.0 ± 21.3
		g_a^-	23.5	290.6 ± 12.7	152.9 ± 23.9	173.1 ± 12.2	268.0 ± 20.3
		g_b^-	9.0	302.3 ± 11.5	255.0 ± 28.8	178.1 ± 11.5	267.5 ± 23.9
Asn in unoccupied sequon	853	g^+	18.5	57.0 ± 16.0	183.8 ± 67.5	NA	NA
		t	39.7	191.9 ± 17.9	182.4 ± 82.5		
		g_a^-	31.5	290.7 ± 14.7	138.3 ± 30.3		
		g_b^-	10.2	289.3 ± 14.3	297.5 ± 41.8		
Surface nonsequon	5200	g^+	15.5	60.1 ± 14.5	177.0 ± 72.0	NA	NA
		t	28.7	190.9 ± 16.5	169.6 ± 83.6		
Asn		g_a^-	38.5	291.5 ± 15.1	134.2 ± 31.3		
		g_b^-	17.3	290.2 ± 17.3	301.2 ± 41.8		
Buried nonsequon	1674	g^+	17.1	61.6 ± 14.5	182.2 ± 72.0	NA	NA
		t	32.4	188.8 ± 16.1	163.1 ± 71.1		
Asn		g_a^-	36.9	291.3 ± 12.6	136.2 ± 34.3		
		g_b^-	13.6	289.3 ± 16.5	300.6 ± 43.2		

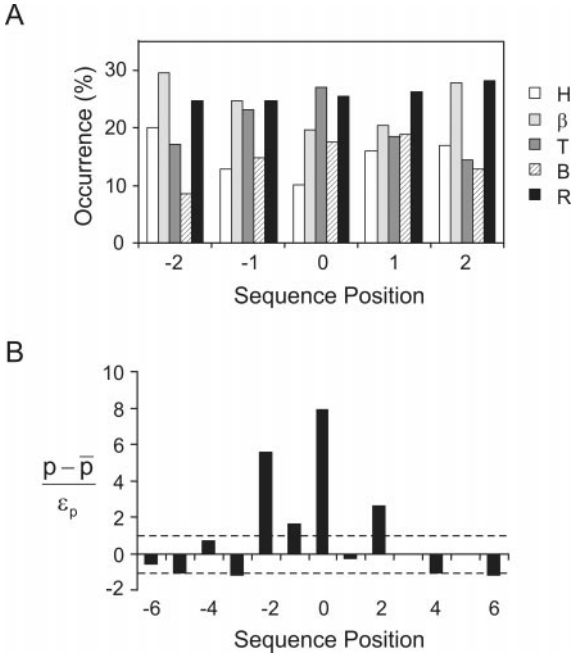

Fig. 4. (**A**) Secondary structure distribution in positions −2 to +2 around occupied glycosylation sites (modified Asn is residue 0). H, helical; β, β-strand; T, turn; B, bend; R, random. (**B**) Variation in the secondary structure change rate in the region −6 to +6 around an occupied glycosylation site, measured in standard deviations relative to the average value. The change rate was measured as the probability of having one type of structure in a given position followed by a different type for the next amino acid. p, probability of a change in secondary structure around an occupied glycosylation site; $\bar{p}$, average probability of a change in secondary structure; ε_p. standard deviation for the average probability of a change in secondary structure (see *Materials and methods* for details).

glycosylation sites is on helices, 10.5% compared to an overall incidence of helical structures of 23%.

Figure 4B shows the probability of a change in secondary structure occurring around a glycosylation site, relative to the average rate of change of secondary structure. It is immediately obvious that the probability of secondary structure changes is strongly related to the presence of a glycan at the glycosylation site having statistically significant peaks in positions −2 and 0 (the glycosylation site), and a smaller increase in position +2.

Surface topology around occupied N*-glycosylation sites*

The surface geometry around all the occupied glycosylation sites in the nonredundant data set have been determined by direct inspection and categorized (see *Materials and methods*). The data are summarized in Table III and specific examples given in Figure 5.

The repertoire of configurations in which glycosylation sites are located is impressive. The majority of sites are found in convex or flat regions of the surface (*xx* and *xf*: 44.2%). There are also a large number of glycosylation sites located on the edge of a groove or cleft in a flat surface (*e*: 18.6%) with the glycan filling partially or totally the depression. Around 10% of the glycans are situated in deep, narrow recesses in the protein surface (*vv*, *vi*), with the first one or two glycan residues in close contact with the surrounding amino acids.

A more quantitative approach to the local surface structure is to measure the accessibility of the Asn side chain atoms to a probe the size of a monosaccharide (~3 Å). As well as a direct measure of the degree of exposure of the Asn residue and hence whether a glycan can be accommodated,

Table III. Analysis of surface geometry and relative accessibility of the Asn residue to a 3 Å probe for all occupied glycosylation sites in the nonredundant data set

Surface Type	Sites (%)	Asn SC (-G)	Asn SC (+G)	Asn Bkb (+G)	Shape
vv	6.0	7.6 ± 5.2	0.0 ± 0.0	0.2 ± 0.6	
vi	3.4	14.9 ± 9.1	1.4 ± 3.6	1.1 ± 3.9	
vx	4.5	21.5 ± 11.1	3.8 ± 7.2	7.6 ± 14.4	
fi	6.5	23.4 ± 10.0	6.3 ± 11.5	3.8 ± 10.6	
ff	6.0	25.1 ± 6.4	3.6 ± 5.6	1.9 ± 5.7	
fv	0.8	29.1 ± 2.7	10.4 ± 10.6	0.0 ± 0.1	
e	18.6	29.5 ± 13.9	13.9 ± 21.9	12.9 ± 23.2	
ii	3.1	32.5 ± 9.9	10.0 ± 9.5	6.0 ± 11.0	
xf	11.0	36.9 ± 11.6	12.6 ± 15.1	6.8 ± 16.3	
xi	6.8	42.4 ± 11.9	16.8 ± 19.7	6.7 ± 11.8	
xx	33.2	57.6 ± 19.8	34.4 ± 27.0	28.2 ± 30.8	

Asn SC (−G), Asn side chain relative accessibility in the absence of the glycan; Asn SC (+G), Asn side chain relative accessibility in the presence of the glycan; Asn Bkb (+G), Asn backbone relative accessibility in the presence of the glycan

the value of the accessibility also gives some indication of the degree of conformational freedom of the glycan core once attached to the Asn. The accessibility values for Asn residues in occupied glycosylation sites are given in Table III. There is a very good correlation between the relative accessibility of the Asn side chain and the topological classification used, the *vv* sites giving the lowest accessibility and the *xx* sites the highest. Table III also gives the relative accessibilities separately for the Asn backbone and side chain atoms, the different geometries showing different patterns of accessibilities.

Figure 6A shows the distribution of relative accessibilities of the unmodified Asn residue in all the potential *N*-glycosylation sites (occupied and unoccupied) in the nonredundant data set. This shows the decrease in accessibility of the Asn residue as a result of the protein tertiary structure, 100% being no decrease. There are surprisingly few potential sites at the very high accessibility end (*xx*) and a large number at the low accessibility end (*vv*). Figure 6B shows the percentage of sites of a given accessibility that are occupied. There is lower level of occupancy (40%) for sites with both very low and very high accessibilities.

The presence of the first GlcNAc residue prevents access to a significant proportion of the protein surface around the glycosylation site (Figure 7). The percentage of the protein surface protected is similar for sites of all accessibilities.

Chemical composition around occupied N-*glycosylation sites*

The average chemical composition of the surface around *N*-glycosylation sites is shown in Figure 8A. As can be seen, there is a significantly above average occurrence of aromatic and Ser/Thr residues and a significantly below average occurrence of basic residues within 5 Å of a glycosylation site. Once the surface concerned is extended to a range of 10 Å from the glycosylation site, the composition moves back toward the average values. The high occurrence of Ser and Thr is not surprising, given their presence in the glycosylation sequon. To filter out such sequence effects, the analysis was repeated excluding all amino acids within the range −2 to +3 of the glycosylation site (results shown as an inset in Figure 8A). The occurrence of aromatic and hydrophobic residues increases somewhat, and there is still a low occurrence of basic residues, but there is now a low occurrence of Ser/Thr as well.

Figure 8B shows the distribution of all amino acid atoms within 4 Å of any glycan atoms, compared to the average composition of the protein surface and the protein core. With the exception of backbone carbonyl oxygen atoms and aromatic side chain atoms, the values for the protein–glycan contact areas lie between those of the surface and the core. The glycan makes proportionally fewer contacts with backbone carbonyl oxygen atoms and proportionally far more contacts with aromatic side chain atoms than would be expected based on the average composition of either the protein surface or the protein core.

Discussion

Due to technical difficulties the number of glycoproteins resolved by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy is very limited. As can be seen from Table I, they represent less than 3% of the total number of reported 3D structures (~17,000), whereas over 50% of the eukaryotic proteins are glycosylated (Apweiler *et al.*, 1999). In the last few years, however, the number of resolved structures has increased, now reaching levels that make feasible a comparative analysis of the properties of the protein around glycosylation sites. There are a number of significant points that emerge from this study.

N-*glycosylation sequon occupancy is modulated by local amino acid sequence but not by remote sequence*

N-glycosylation takes place in the endoplasmic reticulum (ER) and consists of the cotranslational transfer of a $Glc_3Man_9GlcNAc_2$ (G3M9) oligosaccharide from dolichol to an asparagine situated in an Asn-X-Ser/Thr sequon. The process is mediated by the OST complex and its efficiency depends on many cellular factors such as the translation rate, the levels of OST, and the availability of the dolichol bound G3M9. As a result of this complex cellular balance not all the potential glycosylation sites are occupied, the most recent estimate being 66% based on analysis of 749 well-characterized glycoproteins listed in the 75,000 entries in the SWISS-PROT database (Apweiler *et al.*, 1999). The value of 65% occupancy from our survey is very close to this

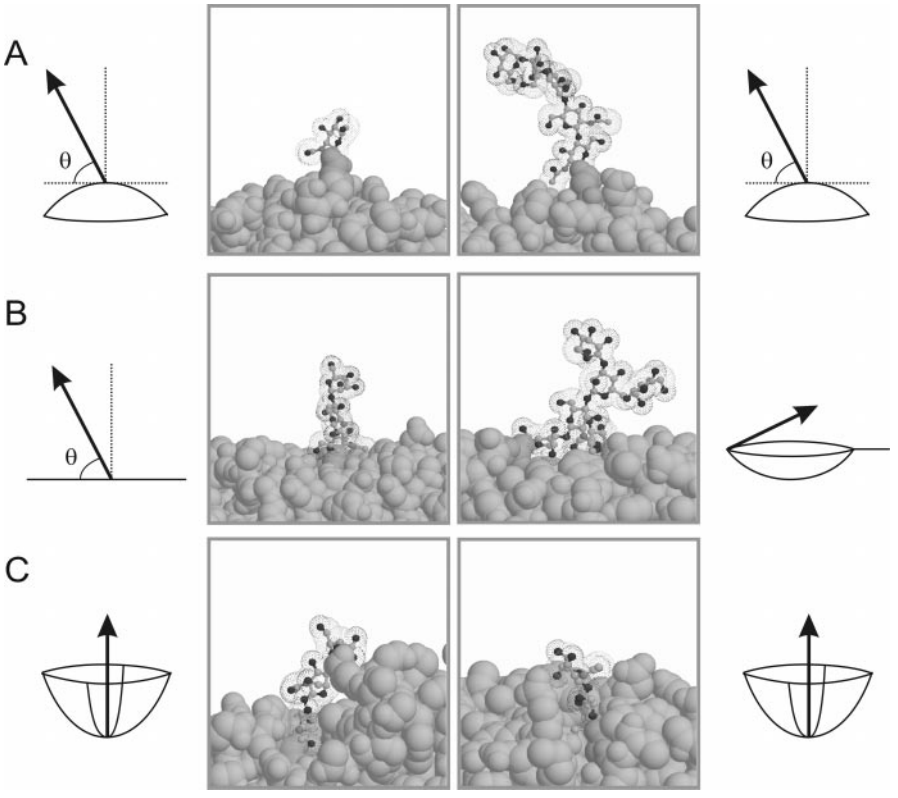

Fig. 5. Examples of protein surface geometry around glycosylation sites, together with schematic representations. (**A**) Sites situated on convex surfaces: left, *xx*, Asn 168 in glucose oxidase, pdb code 1gal (Hecht *et al*., 1993); right, *xx*, Asn 89 in glucose oxidase, pdb code 1gal (Hecht *et al*., 1993). (**B**) Sites in flat environments: left, *ff*, Asn A285 in hemagglutinin, pdb code 1hao (Chen *et al*., 1998); right, *e*, Asn 32 in rhamnogalacturonase A, pdb code 1rmg (Petersen *et al*., 1997). (**C**) Sites situated in deep recesses of the surface: left, *vv*, Asn 301 in acid phosphatase, pdb code 1rpa (Lindqvist *et al*., 1993); right, *vv*, Asn B291 in serine carboxypeptidase II, pdb code 3sc2 (Liao *et al*., 1992).

value and so suggests that most of the "unoccupied" sites are due to absence of glycan rather than unresolved electron density.

At molecular level the process depends on the ability of OST to efficiently recognize *N*-glycosylation sites along the protein chain during translation. When emerging from the translocon the nascent chain may be assumed in an extended configuration, and its recognition by OST is assumed to mainly depend on the amino acids in the close vicinity of the Asn residue to be glycosylated. This is confirmed by our observation that there are no particular patterns of remote amino acid distribution associated with occupied glycosylation sites. However, there are local sequence variations that do appear to affect the level of glycosylation.

There is a large preference for Thr, as opposed to Ser, in position +2. This is in agreement with the observation that replacing Ser with Thr in the sequon results in an overall increase of the occupancy (Kasturi *et al*., 1995). Proline is not present in position +1, and its incidence in position +3 is very low, again consistent with previous studies (Bause, 1983). However, especially in Ser sequons, Pro reaches overexpected levels in positions −6, −3, and +4. A bend of the chain on either side of the sequon may be therefore favorable for recognition by the OST. Only two amino acids, Gly and Val, were found in position +1 with a probability more than one standard deviation above the expected value, in contrast to a previous report that Gly, Asn, and Phe were found preferentially at position +1 (Christlet *et al*., 1999).

There is a clear preference for nonpolar amino acids in positions −2 to +4 and in particular for aromatic residues in position −2 and −1, small nonpolar amino acids in position +1, and bulky hydrophobic amino acids in positions +3 to +5. It is also interesting to note the significant deficit of acidic amino acids upstream of occupied glycosylation sites. This contrasts with the overexpected levels observed for unoccupied sites, suggesting that a negative charge in that region may constitute a handicap for glycosylation. These trends are more accentuated in Ser-type sequons. Thus Asn-X-Ser sequons appear to be less efficiently glycosylated and the glycosylation is more sensitive to other sequence effects, confirming the fragility of their recognition by OST compared to Asn-X-Thr sequons.

Alterations in primary structure could provide a control mechanism for ensuring preferential glycosylation at particular sites. This may prove crucial in conditions that put a strain on the ER glycosylation machinery. There is increasing evidence that not all glycosylation sites are equally important for glycoprotein function. For example, in mouse tyrosinase the absence of glycans at sites Asn 71 and Asn 371 produce loss of activity due to a dramatic

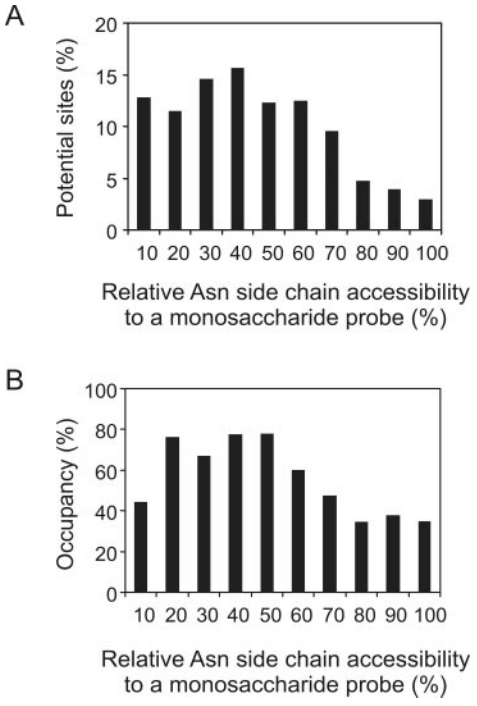

Fig. 6. (**A**) Plot of occurrence (as a percentage of total sites) versus the relative accessibility of the Asn side chain to a 3 Å probe for all potential *N*-glycosylation sites (occupied and unoccupied). For occupied sites, the glycan was removed prior to calculation of the accessibility. Relative accessibility is the accessibility of the Asn side chain divided by the accessibility of the Asn side chain in an extended Gly-Asn-Gly tripeptide. (**B**) Plot of the percentage of sites with a given relative accessibility that are occupied versus the relative accessibility of the Asn side chain to a 3 Å probe.

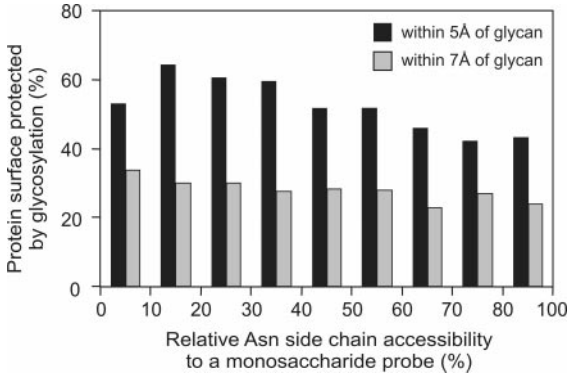

Fig. 7. Plot of the percentage of the protein surface, within a 5 Å or 7 Å radius of the glycosylation site that is made inaccessible to a 1.5 Å probe by the presence of a glycan versus the relative accessibility of the Asn side chain to a 3 Å probe.

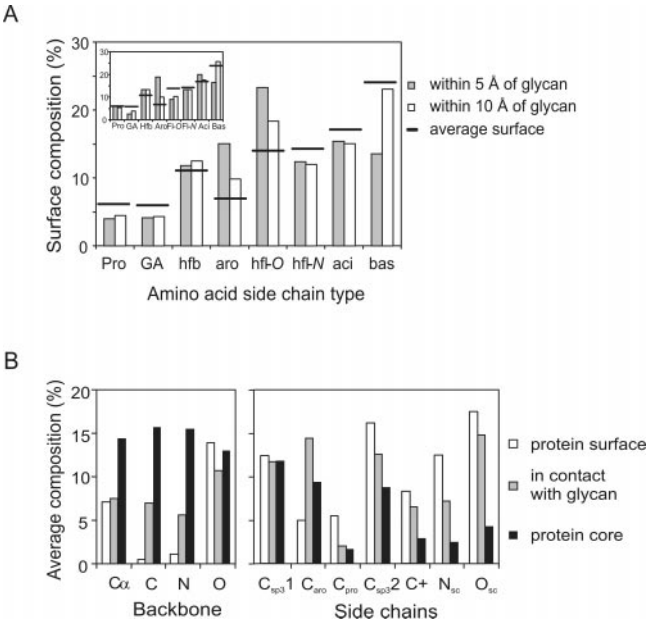

Fig. 8. (**A**) The average composition of the protein surface within 5 Å and 10 Å of *N*-glycosylation sites, compared to the average composition of protein surfaces over the database. *Inset:* The surface composition around occupied glycosylation sites excluding residues from −2 to +3 in the local sequence. Pro, proline; GA, small hydrophobic amino acids (Gly, Ala); hfb, larger hydrophobic amino acids (Val, Leu, Ile, Met); aro, aromatic amino acids (Phe, Tyr, Trp); hfl-O, hydrophilic amino acids with oxygen side chains (Ser, Thr); hfl-N, hydrophilic amino acids with nitrogen side chains (Asn, Gln); aci, acidic amino acids (Asp, Glu); bas, basic amino acids (Lys, Arg, His). (**B**) The distribution of protein atoms within 4 Å of any glycan atom. This is compared to the average atom compositions of the protein surface and of the protein core, measured over the entire database. Cα, C, N, O, peptide backbone atoms using standard pdb nomenclature; $C_{sp3}1$, sp3 carbon atoms in hydrophobic amino acid side chains (Leu, Ile, Val, Ala, Met, Cys, Phe, Trp, Tyr); C_{aro}, sp2 carbon atoms in aromatic rings (Phe, Trp, Tyr); C_{pro}, carbon atoms in proline side chains; $C_{sp3}2$, sp3 carbon atoms in hydrophilic amino acid side chains (Asp, Glu, Asn, Gln, Lys, Arg, His, Thr, Ser); C+, nonsp3 carbon atoms in amino acid side chains which are attached to nitrogen or oxygen (Cγ of Asp and Asn, Cδ of Glu and Gln, Cζ of Arg, ring carbons of His); N_{sc}, side chain nitrogen atoms; O_{sc}, side chain oxygen atoms.

alteration in the protein–calnexin interaction during folding (Branza-Nichita *et al.*, 2000). In contrast, sites Asn 111 and Asn 161 are not occupied at all, while preventing glycosylation at sites Asn 230 and Asn 337 leaves tyrosinase fully functional. Critical sites, such as Asn 71 and Asn 371, need to be occupied under all conditions for function. We note here that indeed Asn 71 and Asn 371 present typical signatures for a preferential recognition by the OST (VFY*N*RTCQC and IFM*N*GTMSQ). It is also interesting to note that hydrophobic amino acids in the vicinity of the glycosylation site appear to increase the affinity of UDP-glucose glycoprotein:glucosyltransferase (UGGT) for short glycopeptide substrates (Taylor *et al.*, 2003).

The Asn-GlcNAc linkage has a single preferred conformation

The Asn-GlcNAc linkage exhibits the same conformation in 90% of the structures. The ψ_N value of 180°, and its relatively narrow distribution is expected for a *trans* amide linkage, only 1.2% of structures showing a *cis* amide linkage. The ϕ_N values occur in a wider range centered at ~240°. These results are in agreement with experimental results on glycopeptides using NMR spectroscopy (Wormald *et al.*,

1991; Davis *et al*., 1994) and a previous statistical analysis of X-ray structures (Imberty and Perez, 1995).

It is interesting to note, however, that from the large data set analyzed here a second minor basin is shaping up. This is centred at ϕ_N ~75°, with a shift of over 30° from the secondary minima of ϕ_N ~40° predicted from energy calculations (Imberty and Perez, 1995). The minor conformer is seen in ~12% of structures and it is well separated from the main ϕ_N conformer.

N-*glycosylation alters the conformational preferences of the Asn residue side chain to more extended conformers and reduces the overall conformational freedom of the side chain*

The conformational preferences of unmodified Asn side chains have been reported to be sensitive to surface exposure (Pickett and Sternberg, 1993). Therefore we looked separately at the side chain conformational distribution for glycosylated Asn residues, unglycosylated Asn residues in a glycosylation sequon, solvent-exposed Asn residues and buried Asn residues. The residues in all four categories exhibited the same three distinct conformations (labelled g^+, g^- and t according to the χ_1 value). Although small differences were observed between the three classes of unmodified Asn residues, their conformational behaviors were broadly similar. However, whereas the unmodified Asn residues showed a distinct preference for the g^- conformer in preference to the t conformer, this was completely reversed for the glycosylated Asn residues. The change from g^- to t has the effect of extending the Asn side chain and increasing the distance between the first GlcNAc residue and the peptide backbone. The rotamer distribution for the modified Asn side chains is similar to that previously reported (Imberty and Perez, 1995), although these authors state that this distribution is the same for unmodified Asn residues.

The second observation is that there is a very significant decrease in the standard deviation of the χ_2 distribution for glycosylated Asn residues compared to the three sets of unmodified Asn residues. This decrease in the conformational space sampled by the Asn side chain on glycosylation is consistent with the proposal that glycans may limit the conformational space available to a peptide backbone (Imperiali and O'Connor, 1999). This restriction may be also interpreted as reduction in the potential dynamic freedom of the Asn side chain on glycosylation. There is extensive experimental evidence that glycosylation results in a short-range decrease in the backbone flexibility of highly flexible peptides (Imperiali and Rickert, 1995; Wormald *et al*., 2002) and a long-range decrease in the backbone flexibility of folded proteins (Wormald and Dwek, 1999). Dynamic studies have recently shown that moving in toward the protein core there is a progressive freezing out of anharmonic dynamics (Dellerue *et al*., 2001). Combining these observations, one could speculate that glycans might alter the dynamics of the Asn side chain by shifting local properties closer to those of a protein core. This would increase the relaxation times of the Asn residue and surrounding side chains and provide a mechanism for the reduction in dynamics to be relayed to more remote parts of the protein.

N-*glycosylation occurs more frequently at points of change in secondary structure*

General dogma suggests that *N*-glycosylation usually occurs on flexible loops. A previous statistical analysis suggested that 25% of glycosylation sites occur on β-turns (Imberty and Perez, 1995). The analysis of the secondary structure around occupied glycosylation sites shows that *N*-glycosylation can occur on all type of secondary structure, with a bias toward turns and bends. However, the most striking observation is that there is a highly increased probability of glycosylation sites occurring at or just after points in the chain where there is a change in secondary structure. This raises the possibility that glycosylation favors reorientation of the peptide chain. As well as stabilizing a particular local fold, glycans may play a role in organizing the folding process by promoting changes in backbone conformation in folding intermediates. The positions of glycosylation sites may have evolved to act as landmarks for ending or starting regions of regular secondary structure to promote efficient folding.

The location of glycosylation sites in regions of low accessibility may be significant for function

In globular proteins, the surface is not smooth, and simple concepts such as local curvature can be used only in a qualitative way. Alternatively one can use accessibility as a quantitative parameter to describe one aspect of the surface geometry. We use both of these approaches to characterize the protein surface around glycosylation sites.

The observed surface topologies around occupied glycosylation sites are very varied. Less than 50% of sites are located on convex surfaces (*xx* and *xf*: 44%). In these cases, the glycan makes either no contacts with the protein surface or very few, usually via the acetamido group of the first GlcNAc residue (for example, see Figure 5A, right). There is a significant number of sites in which the Asn and core glycan residue(s) fill a groove or hole (*vv*, *vi*, and *e*: 28%) and make extensive contacts with the protein surface. The overrepresentation of these sites suggest that such protein–glycan contacts are significant, possibly as a means of stabilizing the local or global structure of the protein as previously mentioned.

Potential glycosylation sites cover the whole range of Asn residue side chain accessibilities from fully accessible to nearly completely buried. The occupancy of sites with different accessibilities is reasonably constant, except for a decreased occupancy at high and very low accessibility. The former has to be treated with some care because this may just reflect the fact that glycans attached to fully accessible Asn residues are likely to be more mobile and so are less likely to be resolved in the electron density. The latter might be expected. However, because protein glycosylation precedes folding the inaccessibility of the Asn in the folded protein does not mean that it is inaccessible to the OST. There are two possible explanations for the absence of a glycan at an inaccessible site: (1) The presence of the glycan leads to protein misfolding and degradation, and so only the small percentage of unglycosylated proteins survive; or (2) the site is not recognized by the OST. Analysis of the amino acid sequence of the subset of unoccupied glycosylation

sites with very low accessibility (less than 2% for a monosaccharide probe) shows that 29% have Pro at position +1 (compare with 12% in the complete set of unoccupied sites), 77% have Ser at position +2 (compare with 59%), and 24% have acidic residues at positions −1 or −2 (compare with 14%). All these factors have been associated with reduced glycosylation site occupancy. Approximately 95% of these inaccessible sites present at least one of these nonoccupation signature elements. This strongly suggests that local primary structure is indeed used as a mechanism for control of occupancy *in vivo*.

Given the hydrophilic nature of monosaccharides, one might have expected there to be fewer occupied glycosylation sites with lower accessibilities and more with higher. The small number of potential sites with a very high accessibility is consistent with the idea that protein–glycan contacts are functionally significant and have been selected for during evolution. The significant number of *N*-glycosylation sites with low accessibilities in folded proteins is interesting from the point of view of the protein folding quality control mechanism because UGGT recognizes the first residue of the glycan core (Parodi, 2000a) and this will be inaccessible in these cases (discussed further shortly).

Glycosylation sites occur on hydrophobic regions of the protein surface

The average chemical composition, determined either by amino acid type or atomic composition, of the protein surface in the region of a glycosylation sites is significantly different from the normal composition of protein surfaces, even after filtering out the effects of the required presence of a Ser or Thr residue in the sequence. In particular, there is a big increase in the presence of aromatic groups, a smaller increase in the presence of other hydrophobic groups, and a significant decrease in the presence of basic groups. The increased occurrence of surface aromatic and hydrophobic groups is still observed after filtering out all the residues that are close in sequence to the glycosylation site. The regions of protein surfaces in contact with glycans show an atomic composition intermediate between that of the average protein core and average protein surface, except for the higher incidence of aromatic side chain atoms. This provides evidence for the hypothesis that in a significant number of cases *N*-linked glycans are involved in covering/stabilizing hydrophobic patches of the protein surface (Toyoda *et al*., 2002) but that stabilization of regions of positive charge by *N*-linked glycans (Wyss *et al*., 1995) is not a general phenomenon.

The role of glycosylation in protein folding

It is becoming apparent that *N*-linked glycans can play a wide variety of roles during protein folding both *in vitro* and *in vivo*. The high incidence of aromatic amino acids distant in sequence but close in space to glycosylation sites, the observation that glycosylation sites occur more frequently at points of change in secondary structure and the large number of glycosylation sites with a low accessibility in the folded protein give rise to several hypotheses regarding folding and folding control mechanisms.

Glycans are thought to directly affect protein folding pathways by reducing the conformational freedom of the local peptide backbone and thus reduce the loss of configurational entropy on folding (Hoffman and Florke, 1998). The analysis of the secondary structure around glycosylation sites also raises the possibility that glycans may act as markers for points where changes in secondary structure occur by stabilizing particular backbone conformations. The increased incidence of aromatic amino acids distant in sequence but close in tertiary fold to glycosylation sites suggests that *N*-linked glycans might have a further, more direct role in folding, by acting as nucleation sites for remote parts of the protein chain rich in aromatic amino acids. This hypothesis is supported by *in vitro* experiments indicating that the *N*-linked glycans promote folding based on glycan–protein hydrophobic interactions (Nishimura *et al*., 1998; Jitsuhara *et al*., 2002).

The recognition and processing of terminal *N*-linked glycan residues in the ER protein folding and quality control mechanisms is well documented (Parodi, 2000a,b). Protein folding requires cycles of reglucosylation by UGGT and calnexin/calreticulin-assisted folding. There is now considerable evidence that the location of the glycosylation sites can be critical. UGGT recognition requires binding to both exposed hydrophobic patches on the surface of the partially folded protein (Sousa and Parodi, 1995), either from residues close in space or close in sequence (Taylor *et al*., 2003) and to the innermost residue of the glycan (Parodi, 2000a). Although there is an increased probability of finding surface hydrophobic amino acid residues in the vicinity of the glycosylation site, this is a short-range effect (up to 5 Å), and by being in such close vicinity to the glycan in the correctly folded protein these residues will evade recognition by UGGT. The observation that many Asn residues in *N*-glycosylation sites have a low accessibility in the folded protein is more intriguing. In these cases, correct local folding will protect the first glycan residue from UGGT recognition and thus provide a direct mechanism of controlling reglucosylation (Parodi, 2000b).

Materials and methods

Glycoprotein and protein crystal structures were obtained from the RCSB PDB (Berman *et al*., 2000). Any protein in which at least one GlcNAc residue attached to an Asn residue could be seen was included, whether or not the rest of the glycan chain or chains at other potential sites could be seen. Structures were rejected if the bond length between the Asn side chain and the GlcNAc residue was too long.

To obtain a nonredundant data set, glycosylation sites were clustered in groups that differ from each other by less than two amino acids in a region between positions −2 to +3 around the central Asn residue (using a longer sequence for comparison resulted in too many structurally similar proteins being classified as different). From each cluster, one entry was selected for the nonredundant data set. The criteria used for selection were a minimal number of unresolved atom positions, a maximal number of glycan units attached to the protein, the accuracy of both protein and glycan stereochemistry, and the resolution. Of the

structures in the nonredundant data set, 9% are of over 3 Å resolution and 25% less than 2 Å resolution. Most of the clusters (~75%) consisted of one to four structures, whereas 11% of the clusters contained more than 10 structures. With very few exceptions, each member of the nonredundant data set had a different protein conformation around the glycosylation site. We note that in few cases two structures in the nonredundant set did have very similar folds, but in this work they were still considered as distinct entities.

The torsion angles used to describe the Asn residue side chain and Asn-GlcNAc linkage conformations are: χ_1 = N–C^{α}–C^{β}–C^{γ}; χ_2 = C^{α}–C^{β}–C^{γ}–N; ψ_N = C^{β}–C^{γ}–N–C1; ϕ_N = C^{γ}–N–C1–O5. Torsion angles are classified as g^+ = +60°, t = 180°, and g^- = −60° (standard IUPAC nomenclature). Histogram plots for each linkage were obtained by counting the number of structures within a specific 10° window (0°–10°, 10°–20°, etc.) (Petrescu *et al.*, 1999). Two conformers were considered distinct if there is at least one distinct minimum between them in the histogram plot of at least one of the torsion angles, adjacent peaks in the histogram are separated by at least 60° and each conformer represented by at least 10% of the sample population. The ranges of torsion angles associated with each conformer can be judged from the width of the peaks in the histogram and the dispersion of the peaks in the torsion angle plots. In cases of doubt we have included rather than excluded structure from the conformer regions, to give the largest possible populations for statistical analysis.

The statistical analysis of the primary structure around each glycosylation site was performed on a 35-amino-acid region between positions −9 and +25. The normal statistical occurrence for each amino acid was determined by measuring their frequencies in each of the polypeptide chains over 100 amino acids long in the nonredundant set of structures (to give the normal occurrence in glycoproteins remote from glycosylation sites, not biased by protein types that are never glycosylated, such as integral membrane proteins). These results were then used to determine an average frequency and standard deviation for each amino acid.

The statistical analysis of the secondary structure around each glycosylation site was performed on the same 35-amino-acid region between positions −9 and +25. The secondary structure for each residue was classified as one of the following: helix (all helices types); extended (all β types); turn (all turns); bend; and random (including β-bridges) using the Kabsch and Sander (1983) method as reported in the structure explorer of the PDB server (www.rscb.org/pdb). The rate of structural change, p, in each position n in the interval (−9, +25) was estimated as the probability of having any given type of secondary structure at n followed by any other type of secondary structure at $n + 1$. The average change rate, $\bar{p}$, and standard deviation, ε_0, were determined using the same 35-amino-acid stretches from the nonredundant data set.

Conventional geometry was used to describe qualitatively by visual inspection the protein surface around the glycosylation site and the orientation of the oligosaccharide. For each site two orthogonal directions were considered, the curvature along these directions evaluated by inspection and the curvature classified as follows: convex (*x*), concave (*v*), inflection (*i*), and flat (*f*). The surface can then be locally defined by the curvature in the two directions. Using this taxonomy, there are 10 types of surfaces: *xx*, *xv*, *xi*, *xf*, *vv*, *vi*, *vf*, *ii*, *if*, and *ff*. We found that very frequently a local *iv*-type surface occurs when the site is situated on the edge of a groove in a larger flat environment. These were labeled separately as *e* (edge). This 11-class taxonomy is represented schematically in Table III and provides a qualitative description of the surface geometry.

Characteristics of the surface around the glycosylation sites were evaluated quantitatively using accessibility as measured by the program NACCESS (Hubbard *et al.*, 1991; Hubbard and Thornton, 1993). The degree of exposure of a given glycosylation site was measured using the accessibility of the Asn residue side chain to a probe of radius 3.0 Å (roughly a monosaccharide) in the absence of any glycans. The relative accessibility was defined as the ratio between the side chain accessibility of the Asn residue and that of an Asn residue located in a fully extended Gly-Asn-Gly tripeptide. The solvent accessibility for the Asn residues in the presence and absence of the glycan was measured by the same method using a probe of radius 1.5 Å.

The chemical composition of the surface around each *N*-glycosylation site was determined by counting the residues accessible to a 1.5 Å probe (size of a water molecule) within a distance of 5 Å or 10 Å from the glycosylation site. This was compared to the average surface composition remote from glycosylation sites determined for the proteins in the nonredundant data set.

Molecular modeling, the inspection of structures and the determination of protein–glycan contacts were carried out on Silicon Graphics Fuel and Octane 2 stations using the programs Insight II and Discover (Accelrys).

Acknowledgments

This work has been supported by the Wellcome Trust CRIG 067361.

Abbreviations

ER, endoplasmic reticulum; NMR, nuclear magnetic resonance; OST, oligosaccharyl transferase complex; PDB, Protein Data Bank; UGGT, UDP-glucose glycoprotein:glucosyltransferase.

References

Apweiler, R., Hermjakob, H., and Sharon, N. (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta*, **1473**, 4–8.

Bause, E. (1983) Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. *Biochem. J.*, **209**, 331–336.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Branza-Nichita, N., Negroiu, G., Petrescu, A.J., Garman, E.F., Platt, F.M., Wormald, M.R., Dwek, R.A., and Petrescu, S.M. (2000) Mutations at critical N-glycosylation sites reduce tyrosinase activity by altering folding and quality control. *J. Biol. Chem.*, **275**, 8169–8175.

Chen, J., Lee, K.H., Steinhauer, D.A., Stevens, D.J., Skehel, J.J., and Wiley, D.C. (1998) Structure of the hemagglutinin precursor cleavage site, a determinant of influenza pathogenicity and the origin of the labile conformation. *Cell*, **95**, 409–417.

Christlet, T.H.T. and Veluraja, K. (2001) Database analysis of O-glycosylation sites in proteins. *Biophys. J.*, **80**, 952–960.

Christlet, T.H.T., Biswas, M., and Veluraja, K. (1999) A database analysis of potential glycosylating Asn-X-Ser/Thr consensus sequences. *Acta Crystallog. D*, **55**, 1414–1420.

Cooper, C.A., Harrison, M.J., Wilkins, M.R., and Packer, N.H. (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.*, **29**, 332–335.

Crocker, P.R. (2002) Siglecs: sialic-acid-binding immunoglobulin-like lectins in cell-cell interactions and signalling. *Curr. Opin. Struct. Biol.*, **12**, 609–615.

Daniels, R., Kurowski, B., Johnson, A.E., and Hebert, D.N. (2003) N-linked glycans direct the cotranslational folding pathway of influenza hemagglutinin. *Mol. Cell*, **11**, 79–90.

Davis, J.T., Hirani, S., Bartlett, C., and Reid, B.R. (1994) ^{1}H NMR studies on an Asn-linked glycopeptide. GlcNAc-1 C2-N2 bond is rigid in H_2O. *J. Biol. Chem.*, **269**, 3331–3338.

Dellerue, S., Petrescu, A.J., Smith, J.C., and Bellissent-Funel, M.C. (2001) Radially softening diffusive motions in a globular protein. *Biophys. J.*, **81**, 1666–1676.

Drickamer, K. and Taylor, M.E. (1993) Biology of animal lectins. *Annu. Rev. Cell Biol.*, **9**, 237–264.

Duus, J.O., Gotfredsen, C.H., and Bock, K. (2000) Carbohydrate structural determination by NMR spectroscopy: modern methods and limitations. *Chem. Rev.*, **100**, 4589–4614.

Dwek, R.A. (1996) Glycobiology: toward understanding the function of sugars. *Chem. Rev.*, **96**, 683–720.

Gagneux, P. and Varki, A. (1999) Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology*, **9**, 747–755.

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

Hecht, H.J., Kalisz, H.M., Hendle, J., Schmid, R.D., and Schomburg, D. (1993) Crystal structure of glucose oxidase from *Aspergillus niger* refined at 2.3 Å resolution. *J. Mol. Biol.*, **229**, 153–172.

Helenius, A. and Aebi, M. (2001) Intracellular functions of N-linked glycans. *Science*, **291**, 2364–2369.

Hoffman, D. and Florke, H. (1998) A structural role for glycosylation: lessons from the hp model. *Folding Design*, **3**, 337–343.

Hubbard, S.J., Campbell, S.F., and Thornton, J.M. (1991) Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J. Mol. Biol.*, **220**, 507–530.

Hubbard, S.J. and Thornton, J.M. (1993) NACCESS [computer software]. Department of Biochemistry and Molecular Biology, University College, London.

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

Imberty, A. and Perez, S. (2000) Structure, conformation and dynamics of bioactive oligosaccharides: theroretical approaches and experimental validations. *Chem. Rev.*, **100**, 4567–4588.

Imperiali, B. and O'Connor, S.E. (1999) Effect of N-linked glycosylation on glycopeptide and glycoprotein structure. *Curr. Opin. Chem. Biol.*, **3**, 643–649.

Imperiali, B. and Rickert, K.W. (1995) Conformational implications of asparagine-linked glycosylation. *Proc. Natl Acad. Sci. USA*, **92**, 97–101.

Jitsuhara, Y., Toyoda, T., Itai, T., and Yamaguchi, H. (2002) Chaperone-like functions of high-mannose type and complex-type N-glycans and their molecular basis. *J. Biochem.*, **132**, 803–811.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kasturi, L., Eshleman, J.R., Wunner, W.H., and Shakin-Eshleman, S.H. (1995) The hydroxy amino acid in an Asn-X-Ser/Thr sequon can influence N-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein. *J. Biol. Chem.*, **270**, 14756–14761.

Lasky, L.A. (1995) Selectin-carbohydrate interactions and the initiation of the inflammatory response. *Annu. Rev. Biochem.*, **64**, 113–139.

Liao, D.I., Breddam, K., Sweet, R.M., Bullock, T., and Remington, S.J. (1992) Refined atomic model of wheat serine carboxypeptidase II at 2.2-Å resolution. *Biochemistry*, **31**, 9796–9812.

Lindqvist, Y., Schneider, G., and Vihko, P. (1993) Three-dimensional structure of rat acid phosphatase in complex with L(+)-tartrate. *J. Biol. Chem.*, **268**, 20744–20746.

Lis, H. and Sharon, N. (1998) Lectins: carbohydrate-specific proteins that mediate cellular recognition. *Chem. Rev.*, **98**, 637–674.

Maly, P., Thall, A., Petryniak, B., Rogers, C.E., Smith, P.L., Marks, R.M., Kelly, R.J., Gersten, K.M., Cheng, G., Saunders, T.L., and others. (1996) The alpha(1,3)fucosyltransferase Fuc-TVII controls leukocyte trafficking through an essential role in L-, E-, and P-selectin ligand biosynthesis. *Cell*, **86**, 643–653.

Nishimura, I., Uchida, M., Inohana, Y., Setoh, K., Daba, K., Nishimura, S., and Yamaguchi, H. (1998) Oxidative refolding of bovine pancreatic RNases A and B promoted by Asn-glycans. *J. Biochem.*, **123**, 516–520.

Parodi, A.J. (2000a) Protein glucosylation and its role in protein folding. *Annu. Rev. Biochem.*, **69**, 69–93.

Parodi, A.J. (2000b) Role of N-oligosaccharide endoplasmic reticulum processing reactions in glycoprotein folding and degradation. *Biochem. J.*, **348**(pt 1), 1–13.

Petersen, T.N., Kauppinen, S., and Larsen, S. (1997) The crystal structure of rhamnogalacturonase A from *Aspergillus aculeatus*: a right-handed parallel beta helix. *Structure*, **5**, 533–544.

Petrescu, A.J., Petrescu, S.M., Dwek, R.A., and Wormald, M.R. (1999) A statistical analysis of N- and O-glycan linkage conformations from crystallographic data. *Glycobiology*, **9**, 343–352.

Petrescu, S.M., Branza-Nichita, N., Negroiu, G., Petrescu, A.J., and Dwek, R.A. (2000) Tyrosinase and glycoprotein folding: roles of chaperones that recognize glycans. *Biochemistry*, **39**, 5229–5237.

Pickett, S.D. and Sternberg, M.J. (1993) Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol.*, **231**, 825–839.

Rudd, P.M. and Dwek, R.A. (1997) Rapid, sensitive sequencing of oligosaccharides from glycoproteins. *Curr. Opin. Biotechnol.*, **8**, 488–497.

Rudd, P.M., Wormald, M.R., Stanfield, R.L., Huang, M., Mattsson, N., Speir, J.A., DiGennaro, J.A., Fetrow, J.S., Dwek, R.A., and Wilson, I.A. (1999) Roles for glycosylation of cell surface receptors involved in cellular immune recognition. *J. Mol. Biol.*, **293**, 351–366.

Sousa, M. and Parodi, A.J. (1995) The molecular basis for the recognition of misfolded glycoproteins by the UDP-Glc:glycoprotein glucosyltransferase. *Eur. Mol. Biol. Org. J.*, **14**, 4196–4203.

Taylor, S.C., Thibault, P., Tessier, D.C., Bergeron, J.J., and Thomas, D.Y. (2003) Glycopeptide specificity of the secretory protein folding sensor UDP-glucose glycoprotein:glucosyltransferase. *EMBO Reports*, **4**, 405–411.

Toyoda, T., Arakawa, T., and Yamaguchi, H. (2002) N-glycans stabilize human erythropoietin through hydrophobic interactions with the hydrophobic protein surface: studies by surface plasmon resonance analysis. *J. Biochem.*, **131**, 511–515.

Varki, A., Cummings, R., Esko, J., Freeze, H., Hart, G., and Marth, J. (1999) *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Wormald, M.R. and Dwek, R.A. (1999) Glycoproteins—glycan presentation and protein stability. *Structure*, **7**, R155–R160.

Wormald, M.R., Petrescu, A.J., Pao, Y.L., Glithero, A., Elliott, T., and Dwek, R.A. (2002) Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modelling. *Chem. Rev.*, **102**, 371–386.

Wormald, M.R., Wooten, E.W., Bazzo, R., Edge, C.J., Feinstein, A., Rademacher, T.W., and Dwek, R.A. (1991) The conformational effects of N-glycosylation on the tailpiece from serum IgM. *Eur. J. Biochem.*, **198**, 131–139.

Wyss, D.F., Choi, J.S., Li, J., Knoppers, M.H., Willis, K.J., Arulanandam, A.R., Smolyar, A., Reinherz, E.L., and Wagner, G. (1995) Conformation and function of the N-linked glycan in the adhesion domain of human CD2. *Science*, **269**, 1273–1278.