Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A database analysis of potential glycosylating Asn-*X*-Ser/Thr consensus sequences

T. Hema Thanka Christlet,[a] M. Biswas[b] and K. Veluraja[a]*

[a]Department of Physics, Manonmaniam Sundaranar University, Tirunelveli 627 012, Tamil Nadu, India, and [b]Bioinformatics Centre, Indian Institute of Science, Bangalore 560 012, India

Correspondence e-mail: bio@md3.vsnl.net.in

Received 13 January 1999
Accepted 28 April 1999

An analysis of the frequency of occurrence of various residues at position X was carried out on the consensus glycosylating sequence Asn-X-Ser/Thr using the PDB three-dimensional database. 488 non-homologous proteins bearing 696 Asn-X-Ser/Thr ($X \neq$ Pro) sequences were analysed. More than 65% of Asn residues, when they occur as part of the consensus sequence, lie on the surface of the protein, implying a potentiality for glycosylation. A deviation parameter (DP) was calculated as a measure of preferential (positive) or non-preferential (negative) selection. At the X position in the consensus-sequence segment, the amino acids Gly, Asn and Phe have statistically significant positive DP values. The high value of DP for Asn is a consequence of the preferential occurrence of homodoublets, while for Phe it may be a consequence of the stacking interaction of the aromatic ring with the glycan. Gly at the X position in the consensus glycosylating sequence may be functionally significant owing to its preference and its high percentage of occurrence in proteins. The Ramachandran (Φ,Ψ) angles around Gly in the consensus sequence show clustering in the region which is disallowed for non-glycyl residues. In this region, a hydrogen bond between the side chain of Asn and the peptide backbone/side chain of Ser/Thr is possible, reflecting a positional as well as a conformational role in the consensus glycosylating sequence. For the 44 confirmed N-glycosylating sequences, an in-depth analysis of the (Ψ^N, Φ^X, Ψ^X, $\Phi^{S/T}$) dihedral angles, which position the side chains of Asn and Ser/Thr, shows that these can be grouped into nine conformational states. In most cases, a direct or water-mediated hydrogen bond between OD1 of Asn and OG of Ser/Thr is possible, reflecting the possible importance of this hydrogen bonding in the glycosylation process.

1. Introduction

Glycosylation is a protein-modification reaction. It generates recognition structures for interaction with external ligands and there is an intimate relation between glycosylation and protein folding (Elbein, 1991; Li *et al.*, 1993; Holst *et al.*, 1996; Gahmberg & Tolvanen, 1996). In N-linked glycosylation the carbohydrate moiety is linked to the amino N atom of the side chain of Asn, whereas in O-glycosylation carbohydrates are linked to the hydroxyl O atom of the side chain of Ser or Thr. A single glycoprotein may contain N-glycosides as well as O-glycosides. N-glycans play an important role in intracellular recognition, positioning of antigens on the cell surfaces, immunoglobulin secretion, lysosomal targeting, immunoregulation *etc.* and have hence attracted special attention (see review by Avanov, 1991).

The N-glycosylation site comprises an Asn-*X*-Ser/Thr tripeptide sequence (Marshall, 1972), where *X* can be any amino acid except Pro. The glycan is transferred by an oligosaccharyltransferase to the Asn of the consensus sequence. N-glycosylation is a co-translational process mediated by the hydrogen-bonding interaction between the GlcNAc residue and Asn (Hubbard & Ivatt, 1981; Abbadi *et al.*, 1986; Imperiali & Shannon, 1991). It has been shown that the affinity of the oligosaccharyltransferase increases with the length of the acceptor peptide and with the amino acid at position *X* (Ronin *et al.*, 1978). A minimum chain length of a pentapeptide is required for N-glycosylation (Bause & Hettkamp, 1979). Peptides in which Asn is at the free N-terminal are not glycosylated. Inhibition studies and statistical analysis have revealed that Pro in the *X* position prevents glycosylation (Bause, 1983; Roitsch & Lehle, 1989; Gavel & von Heijine, 1990). Many workers have analyzed the consensus sequence, and the results show that the *X* position is preferentially occupied by non-bulky amino acids such as Gly (Mononen & Karjalainen, 1984; Gavel & von Heijine, 1990; Imberty & Perez, 1995). Subsequent studies have also shown that glycosylation increases when *X* is Gly (Bause & Hettkamp, 1979), indicating the importance of Gly at the *X* position. In proteins, not all Asn-*X*-Ser/Thr consensus sequences are glycosylated (Bush, 1982; Abbadi *et al.*, 1986).

A detailed analysis of the available consensus and confirmed N-glycosylating sequences in proteins of known three-dimensional structure has been undertaken in an attempt to throw more light on the N-glycosylation process, which plays a dominant role in molecular recognition.

2. Method of calculation

2.1. Data set

The protein structures used in this analysis were taken from the Protein Data Bank using the PDB_SELECT subdatabase (November 1996; Hobohm *et al.*, 1992). Proteins used for analysis had less than 25% sequence identity. A total of 488 proteins, containing 130661 amino acids and bearing 696 Asn-*X*-Ser/Thr consensus glycosylating sequences in which $X \neq$ Pro and Asn is not at the free N-terminal, were used for the analysis. The selected proteins are shown in Fig. 1.

2.2. Identification of spatial neighbours of Asn

Table 1
Percentage of Asn with various spatial neighbours in the Asn-*X*-Ser/Thr consensus sequence.

Number of spatial neighbours	Percentage of Asn
0	13.2
1	11.1
2	22.1
3	18.7
4	12.9
5	12.1
6	6.7
7	2.2
8	0.9
9	0.1

To separate those Asn in the consensus sequence which are likely to be on the surface of the protein from those which are buried, a method based on the number of spatial neighbours of Asn was followed (Panjikar *et al.*, 1997). C$^\alpha$ atoms falling within a sphere of radius 6.5 Å around each Asn (C_i^α) along the chain, excluding the four sequential neighbours C_{i+1}^α, C_{i+2}^α, C_{i-1}^α and C_{i-2}^α of C_i^α, were identified as spatial neighbours. For proteins with more than one identical subunit, only one was considered for computation. Asn residues with less than four spatial neighbours were considered to be potential glycosylating sites, since these are the residues most likely to be on the surface of the protein. The sequences containing Asn residues satisfying the above criteria were grouped as set I and the remaining sequences, in which Asn had more than three spatial neighbours, were grouped as set II.

2.3. Deviation parameter (DP)

The expected frequency of occurrence of each amino-acid residue in the data set (Fig. 1) was calculated as

$$P_{\rm expected}(A) = \textstyle\sum N_i(A) / \sum T_i,$$

where $N_i(A)$ is the number of residues of type A in protein i and T_i is the total number of amino acids in the protein i, where i ranges from 1 to n, n being the total number of proteins.

The observed count $P_{\rm observed}(A)$ was calculated by counting the number of times each amino acid occurred at position X for all the sequences in set I and set II.

$$P_{\rm observed}(A) = N_X(A)/m,$$

where $N_X(A)$ is the number of amino-acid residues of type A at position X in the consensus sequences of set I or set II and m is the number of consensus sequences in set I or in set II.

The deviation parameter (DP), expressed as a percentage, for each amino acid at position X was calculated as

$$\mathrm{DP}(A) = 100\,[P_{\rm observed}(A) - P_{\rm expected}(A)]/P_{\rm expected}(A).$$

A positive value for DP is an indication of preferential occurrence and a negative value is an indication of non-preferential occurrence.

The frequency of occurrence follows counting statistics and hence is associated with an error proportional to the square root of the observed count.

$$\text{Error in counting statistics } \sigma = [P_{\rm observed}(A)]^{1/2}.$$

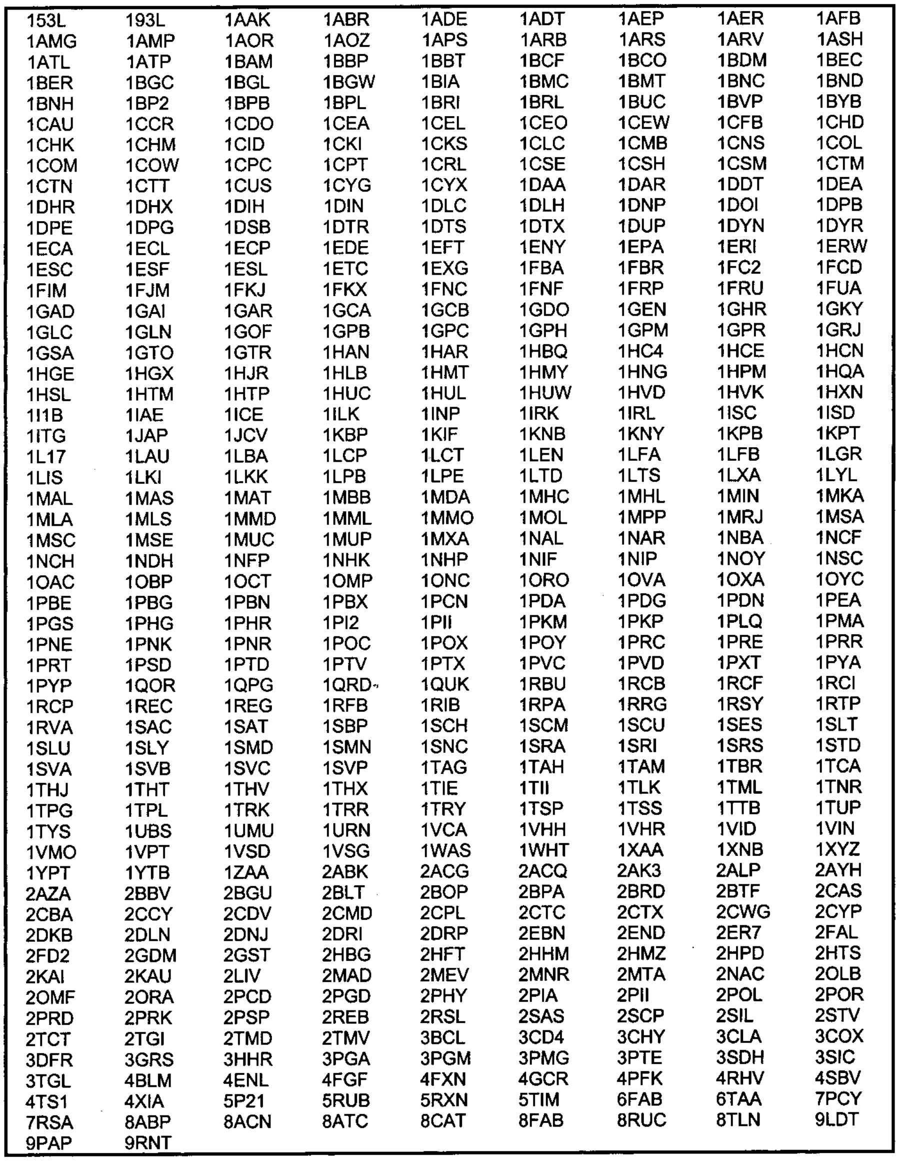
If the difference between the actual count and the observed count for a particular amino acid is equal to or greater than twice the error level (2σ), then the DP was assumed to be statistically significant.

2.4. Ramachandran angle calculation

It has been suggested that peptide conformation might signal glycosylation (Davis *et al.*, 1994). In many cases, the glycosylated tripeptide forms a β-turn (Bush, 1982; Bause, 1983; Imperiali & Rickert, 1995). The Ramachandran (Ψ^N,Φ^X, Ψ^X, $\Phi^{S/T}$) angles which dictate the orientation of the side chain of Asn and Ser/Thr of the signal peptide are shown in Fig. 2. These angles were computed for the 44 confirmed N-glycosylating sequences reported by Imberty & Perez (1995), in order to obtain insight into the conformational aspects of the glycosylating sequence. Ramachandran angles were also calculated for the set I and set II glycosylating tripeptides containing Gly at the *X* position.

3. Results and discussion

Using the protein data set (Fig. 1), the number of spatial neighbours for Asn in the consensus sequences were calculated (Table 1). About 65% of Asn residues in the consensus sequences have less than four neighbours. These sequences are likely to be on the surface of the protein and were grouped in set I. The remaining Asn residues with four or more spatial neighbours were grouped as set II. These are the buried sequences. An examination of 44 confirmed glycosylating sequences reported by Imberty & Perez (1995), shows that more than 85% of the glycosylated Asn have three or fewer spatial neighbours. The Pearson correlation coefficient calculated for variation of the amino-acid residue at *X* in set I and set II was 0.5, indicating poor correlation of the amino-acid residues at position *X* in the two sets. This suggests a difference in preference for different residues at position *X* in set I and set II.



153L	193L	1AAK	1ABR	1ADE	1ADT	1AEP	1AER	1AFB
1AMG	1AMP	1AOR	1AOZ	1APS	1ARB	1ARS	1ARV	1ASH
1ATL	1ATP	1BAM	1BBP	1BBT	1BCF	1BCO	1BDM	1BEC
1BER	1BGC	1BGL	1BGW	1BIA	1BMC	1BMT	1BNC	1BND
1BNH	1BP2	1BPB	1BPL	1BRI	1BRL	1BUC	1BVP	1BYB
1CAU	1CCR	1CDO	1CEA	1CEL	1CEO	1CEW	1CFB	1CHD
1CHK	1CHM	1CID	1CKI	1CKS	1CLC	1CMB	1CNS	1COL
1COM	1COW	1CPC	1CPT	1CRL	1CSE	1CSH	1CSM	1CTM
1CTN	1CTT	1CUS	1CYG	1CYX	1DAA	1DAR	1DDT	1DEA
1DHR	1DHX	1DIH	1DIN	1DLC	1DLH	1DNP	1DOI	1DPB
1DPE	1DPG	1DSB	1DTR	1DTS	1DTX	1DUP	1DYN	1DYR
1ECA	1ECL	1ECP	1EDE	1EFT	1ENY	1EPA	1ERI	1ERW
1ESC	1ESF	1ESL	1ETC	1EXG	1FBA	1FBR	1FC2	1FCD
1FIM	1FJM	1FKJ	1FKX	1FNC	1FNF	1FRP	1FRU	1FUA
1GAD	1GAI	1GAR	1GCA	1GCB	1GDO	1GEN	1GHR	1GKY
1GLC	1GLN	1GOF	1GPB	1GPC	1GPH	1GPM	1GPR	1GRJ
1GSA	1GTO	1GTR	1HAN	1HAR	1HBQ	1HC4	1HCE	1HCN
1HGE	1HGX	1HJR	1HLB	1HMT	1HMY	1HNG	1HPM	1HQA
1HSL	1HTM	1HTP	1HUC	1HUL	1HUW	1HVD	1HVK	1HXN
1I1B	1IAE	1ICE	1ILK	1INP	1IRK	1IRL	1ISC	1ISD
1ITG	1JAP	1JCV	1KBP	1KIF	1KNB	1KNY	1KPB	1KPT
1L17	1LAU	1LBA	1LCP	1LCT	1LEN	1LFA	1LFB	1LGR
1LIS	1LKI	1LKK	1LPB	1LPE	1LTD	1LTS	1LXA	1LYL
1MAL	1MAS	1MAT	1MBB	1MDA	1MHC	1MHL	1MIN	1MKA
1MLA	1MLS	1MMD	1MML	1MMO	1MOL	1MPP	1MRJ	1MSA
1MSC	1MSE	1MUC	1MUP	1MXA	1NAL	1NAR	1NBA	1NCF
1NCH	1NDH	1NFP	1NHK	1NHP	1NIF	1NIP	1NOY	1NSC
1OAC	1OBP	1OCT	1OMP	1ONC	1ORO	1OVA	1OXA	1OYC
1PBE	1PBG	1PBN	1PBX	1PCN	1PDA	1PDG	1PDN	1PEA
1PGS	1PHG	1PHR	1PI2	1PII	1PKM	1PKP	1PLQ	1PMA
1PNE	1PNK	1PNR	1POC	1POX	1POY	1PRC	1PRE	1PRR
1PRT	1PSD	1PTD	1PTV	1PTX	1PVC	1PVD	1PXT	1PYA
1PYP	1QOR	1QPG	1QRD	1QUK	1RBU	1RCB	1RCF	1RCI
1RCP	1REC	1REG	1RFB	1RIB	1RPA	1RRG	1RSY	1RTP
1RVA	1SAC	1SAT	1SBP	1SCH	1SCM	1SCU	1SES	1SLT
1SLU	1SLY	1SMD	1SMN	1SNC	1SRA	1SRI	1SRS	1STD
1SVA	1SVB	1SVC	1SVP	1TAG	1TAH	1TAM	1TBR	1TCA
1THJ	1THT	1THV	1THX	1TIE	1TII	1TLK	1TML	1TNR
1TPG	1TPL	1TRK	1TRR	1TRY	1TSP	1TSS	1TTB	1TUP
1TYS	1UBS	1UMU	1URN	1VCA	1VHH	1VHR	1VID	1VIN
1VMO	1VPT	1VSD	1VSG	1WAS	1WHT	1XAA	1XNB	1XYZ
1YPT	1YTB	1ZAA	2ABK	2ACG	2ACQ	2AK3	2ALP	2AYH
2AZA	2BBV	2BGU	2BLT	2BOP	2BPA	2BRD	2BTF	2CAS
2CBA	2CCY	2CDV	2CMD	2CPL	2CTC	2CTX	2CWG	2CYP
2DKB	2DLN	2DNJ	2DRI	2DRP	2EBN	2END	2ER7	2FAL
2FD2	2GDM	2GST	2HBG	2HFT	2HHM	2HMZ	2HPD	2HTS
2KAI	2KAU	2LIV	2MAD	2MEV	2MNR	2MTA	2NAC	2OLB
2OMF	2ORA	2PCD	2PGD	2PHY	2PIA	2PII	2POL	2POR
2PRD	2PRK	2PSP	2REB	2RSL	2SAS	2SCP	2SIL	2STV
2TCT	2TGI	2TMD	2TMV	3BCL	3CD4	3CHY	3CLA	3COX
3DFR	3GRS	3HHR	3PGA	3PGM	3PMG	3PTE	3SDH	3SIC
3TGL	4BLM	4ENL	4FGF	4FXN	4GCR	4PFK	4RHV	4SBV
4TS1	4XIA	5P21	5RUB	5RXN	5TIM	6FAB	6TAA	7PCY
7RSA	8ABP	8ACN	8ATC	8CAT	8FAB	8RUC	8TLN	9LDT
9PAP	9RNT							

Figure 1
Proteins used in the data set.

The deviation parameters (DP) for the amino acids at position *X* of the consensus sequences in set I and set II are given in Tables 2 and 3, respectively. The DP for Pro is not given, since sequences containing Pro at position *X* were eliminated from the analysis. It is seen from Table 2 that in set I sequences, the amino acids Gly, Asn, Phe and Trp have high positive DP values. The difference between the observed count and the expected count is greater than 2σ for Gly and close to 2σ for Asn, indicating that the preference for Gly and Asn at the *X* position might be significant. For Phe, the difference between the two counts is close to the 1.5σ level, indicating a preferential trend. The high DP for Phe probably reflects the plausible stacking interaction between the aromatic ring of this residue with the glycan, as suggested previously (Imberty & Perez, 1995). The positive DP for Asn at position *X* may be an indication of the preferential occurrence of homodoublets observed previously (Veluraja & Mugilan, 1997). The amino acid Gly, which also has a high positive DP at the *X* position in the consensus sequences of set I, also occurs frequently in proteins. In set II sequences, these amino acids do not have high DP values: Gly has a significant (2σ) negative DP value (Table 3), indicating suppression of Gly at position *X* in the buried sequences. An analysis of the amino-acid sequences from the SWISS-PROT database carried out by Veluraja & Mugilan (1997) revealed that the triplet sequences Asn-Gly-Thr and Asn-Gly-Ser occur more frequently than expected. Thus, the conformational features of Gly in the consensus sequences of set I may be of importance in understanding the nature of the recognition site for glycosylation. Interestingly, Val and Ile are preferred at the *X* position in the buried

Table 2
Deviation parameter (DP) for amino acids at position X in the consensus sequence in set I data.

Amino acid	Observed count	Expected count	Difference in count	Error σ	DP(A)
Ala	31	38.1	7.1	5.6	−18.5
Leu	39	38.1	0.9	6.2	2.5
Gly	52	35.8	16.2	7.2	45.4
Val	33	31.4	1.6	5.8	5.0
Ser	19	27.9	8.9	4.4	−31.9
Glu	25	27.6	2.6	5.0	−9.5
Asp	23	26.9	3.9	4.8	−14.6
Thr	26	26.9	0.9	5.1	−3.2
Lys	25	26.2	1.2	5.0	−4.7
Ile	27	25.2	1.8	5.2	7.1
Asn	32	21.4	10.6	5.7	49.2
Arg	15	21.2	6.2	3.9	−29.1
Phe	27	18.4	8.6	5.2	46.7
Gln	20	17.2	2.8	4.5	16.5
Tyr	20	16.8	3.2	4.5	19.0
His	12	10.1	1.9	3.5	18.3
Met	8	9.5	1.5	2.8	−16.1
Trp	11	6.8	4.2	3.3	61.6
Cys	8	6.4	1.6	2.8	25.7

Table 3
Deviation parameter (DP) for amino acids at position X in the consensus sequence in set II data.

Amino acid	Observed count	Expected count	Difference in count	Error σ	DP(A)
Ala	20	20.4	0.4	4.5	−2.0
Leu	22	20.4	1.6	4.7	7.8
Gly	12	19.2	7.2	3.5	−37.5
Val	27	16.9	10.1	5.2	60.2
Ser	14	15.0	1.0	3.7	−6.5
Glu	9	14.8	5.8	3.0	−39.2
Asp	3	14.4	11.4	1.7	−79.2
Thr	18	14.4	3.6	4.2	24.9
Lys	10	14.1	4.1	3.2	−28.9
Ile	22	13.5	8.5	4.7	62.7
Asn	13	11.5	1.5	3.6	13.0
Arg	19	11.4	7.6	4.4	67.4
Phe	13	9.9	3.1	3.6	31.7
Gln	6	9.2	3.2	2.5	−34.9
Tyr	10	9.0	1.0	3.2	10.9
His	7	5.4	1.6	2.7	28.6
Met	6	5.1	0.9	2.5	17.4
Trp	6	3.7	2.3	2.5	64.4
Cys	6	3.4	2.6	2.5	75.7

sequences (set II) because of their hydrophobicity.

To investigate this further, we extended our analysis to the conformational features of Gly in the consensus sequences. The Ramachandran (Φ^G, Ψ^G) angles were calculated at Gly in each of the 52 available Gly-containing consensus sequences from set I. These Gly residues frequently occur in the conformational space which is disallowed for non-glycyl residues (Fig. 3). The figure shows that there is a clustering in the region $\Phi^G \simeq +60$ to $+110^\circ$ and $\Psi^G \simeq -30$ to $+30^\circ$. 50% of the Gly residues in the consensus sequences from set I occur in this region. Further conformational analysis on the eight confirmed Gly-containing glycosylated sequences revealed that seven of the (Φ^G, Ψ^G) angles fall in this region (Fig. 3). A similar analysis on the 12 available Gly-containing sequences from set II shows that only three (25%) of the (Φ^G, Ψ^G) angles fall in this region.

It has been proposed by earlier workers that the catalytic function of the glycosyltransferases requires a hydrogen-bonded interaction between the side chain of Asn and the hydroxy amino acid (Bause & Legler, 1981). The peptide backbone was shown to adopt a special Asx-turn conformation (Abbadi *et al.*, 1991; Imperiali *et al.*, 1992). Some of these conformations fall in the specified clustering region. An analysis of possible hydrogen bonds based on the criteria of hydrogen-donor to hydrogen-acceptor distance of 2.4–3.6 Å was carried out for the 453 available Asn-X-Ser/Thr sequences in set I. It was found that 33 sequences show a possible hydrogen bond between OD1 of the Asn side chain and the hydroxyl group of the Ser/Thr side chain (OG), 46 sequences have a possible hydrogen bond between the OD1 of the Asn side chain and the N—H of the peptide backbone of the Ser/Thr, and 18 sequences show a possible hydrogen bond between ND2 of the Asn side chain and the hydroxyl group of Ser/Thr. In some sequences more than one hydrogen bond was found, whilst in others none were found. In the sequences which do not exhibit the hydrogen-bonding pattern, the distance between OD1 and OG is in the range 3.7–11 Å, indicating that there may be a possibility of water-mediated hydrogen bonding between these two groups. It is known that water interconnects the side chains of amino acids by linking them through hydrogen bonds. Water-mediated hydrogen bonds also stabilize the structure in carbohydrates (Veluraja & Atkins, 1987). Hence, attempts were made to find potential water-mediated hydrogen bonds in the consensus sequences which do not have direct hydrogen bonds. It was found that 11 sequences exhibit water-mediated hydrogen bonds between OD1 of Asn and the hydroxyl O atom of the side chain of Ser/Thr.

An in-depth analysis on the 52 Asn-Gly-Ser/Thr available sequences in set I shows that nine sequences show a possible hydrogen bond between OD1 and OG, ten sequences have a possible hydrogen bond between OD1 of the Asn side chain and the N—H of the peptide backbone of the Ser/Thr, and six sequences show a possible hydrogen bond between ND2 of the Asn side chain and OG of Ser/Thr. It is interesting to note that in the sequences where hydrogen bonding is possible, the

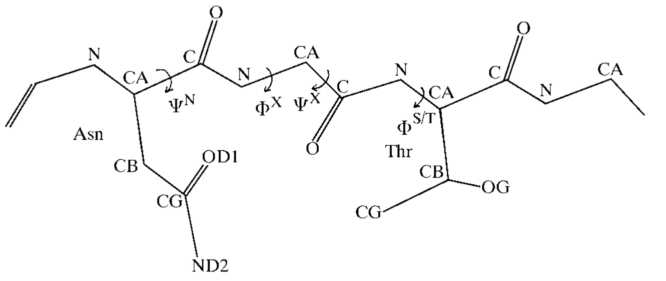


Figure 2
Schematic representation of the Asn-X-Ser/Thr peptide fragment along with the dihedral angles Ψ^N, Φ^X, Ψ^X, $\Phi^{S/T}$ which fix the mutual orientation of the side chains of Asn and Ser/Thr.

Table 4
Direct and water-mediated hydrogen bonds between OD1 and OG observed in proteins with confirmed N-glycosylating sequences.

The coordinates of water are given if there is a possibility of forming water-mediated hydrogen bonds.

Group number	Ψ^N, Φ^X, Ψ^X, $\Phi^{S/T}$ (°) with deviation of ±30°	Protein code	Glycosylation site	OD1–OG distance (Å)	Hydrogen-bonding pattern†	Coordinates of water OW	$\chi^{S/T}$‡ (°)
I	−20, 75, 0, −90	1hge	N154B	2.7	d		
		1lyb	N70	2.9	d		
		2bat	N86	2.9	d		
		2ren	N75	2.3	d		
		1gly	N395	3.5	d		−110 (−117)
		2aai	N95	3.5	d		−60 (−68)
		2bat	N234	7.6	w	89.5, 90.4, 34.8	0 (−50)
II	85, −90, −30, −60	1lga	N257	3.3	d		
		1ova	N298	3.3	d		
		2dnj	N18	3.6	d		
		1gal	N388	3.5	d		−30 (−39)
		1cel	N270	3.5	d		0 (−69)
		3sc2	N291	5.5	w	59.1, 44.9, 94.6	
III	30, −100, 130, −110	2aai	N135	7.2	w	29.3, 48.2, 27.1	−70 (−89)
		2ach	N104	6.7	w	43.0, 73.7, 60.7	
		2bat	N200	8.5	w	104.4, 79.0, 57.0	−110 (151)
		1gal	N89	9.1			
		1gly	N171	10.0			
		1hge	N81	9.8			
		1lfi	N137	9.9			
IV	90, −110, 140, −120	1hge	N165	7.9	w	17.2, 85.5, 18.5	−80 (−170)
		1tmt	N60G	7.9	w	83.8, 21.8, 51.8	−110 (91)
		2fbj	N156	3.9	w	−10.8, 6.0, 15.3	
		1ppf	N159	9.3			
		1aoz	N92	9.8			
V	160, −100, 130, −90	1hge	N38	7.2	w	34.4, 19.6, 35.5	−120 (−175)
		2phl	N228	6.4	w	10.2, 4.5, 73.6	
		1lyb	N199	9.9			
		1tca	N74	8.7			
		2fbj	N59	10.3			
VI	−25, −100, 110, −90	1arp	N143	6.6	w	18.5, 23.2, 52.5	
		1lte	N17	9.8			
		1thg	N364	8.1			
		3sc2	N105	8.9			
VII	−115, −100, −40, −45	2spt	N101	5.0	w	10.7, 55.5, 105.0	−50 (−111)
		2ach	N70	7.6			
VIII	−25, −100, −50, −90	1gal	N355	8.7			
		1lfi	N478	9.4			
		1nsc	N283	9.8			
		1ppf	N109	9.6			
		1thg	N283	7.3			
		3sc2	N113	9.1			
IX	130, 70, 25, −85	1led	N18	7.6			
		2bat	N146	7.0			

† d, direct hydrogen bonding; w, water-mediated hydrogen bonding. ‡ Crystallographic orientation is given in parentheses.

backbone conformation of Gly tends to occur in the specified clustering region. A detailed analysis of the backbone conformational angles for which hydrogen-bond formation was possible revealed that the backbone dihedral angles Ψ^N, Φ^G, Ψ^G and $\Phi^{S/T}$ take values around −20, 85, 0 and −100°, respectively, with deviations of ±30°. This indicates that the backbone adopts a similar conformation in Gly-containing consensus sequences and that hydrogen-bond formation is possible. In these sequences, the side chains of Asn and Ser/Thr may adopt two conformational orientations (−135 ± 45 or +135 ± 45°). For the few sequences for which the (Φ^G, Ψ^G) angles fall in the clustered region for which hydrogen-bond formation is not possible, the backbone conformation differs mainly in the value of Ψ^N, which is around 25 ± 30°. For the Gly containing consensus sequences of set II, hydrogen-bond formation is unlikely, indicating the possible importance of hydrogen bonding in glycosylation. Thus, the presence of Gly at the *X* position allows the backbone to take up a conformation which would normally be disallowed for other amino acids at this position, as seen in Fig. 3. These conformations facilitate the formation of hydrogen bonds between the Asn side chain and the Ser/Thr side chain.

In order to gain further insight into the conformational aspect of the 44 confirmed sequences for N-glycosylation, the Ramachandran angles (Ψ^N, Φ^X, Ψ^X, $\Phi^{S/T}$) which are responsible for orientating the side chains of Asn and Ser/Thr were computed. Analysis of these dihedral angles shows that these can be grouped into nine conformational states, with the marginal deviation of ±30°. These conformational states are listed in Table 4, along with the PDB file name and the site of glycosylation. In group I, the amino acid favoured at the *X* position is Gly, with the exception of 2bat where it is Trp. In this conformational group, Gly falls in the clustered region (Fig. 3). In most of these conformations, a direct hydrogen bond between OD1 and OG is possible, as reported by Imberty & Perez (1995), with the exceptions of 1gly and 2aai. However, a direct hydrogen bond can be facilitated in 1gly and 2aai by a small change in the side-chain dihedral angle of Ser/Thr, as indicated in Table 4. The group II conformational state favours the allowed values (85, −90, −30, −60°) of the Ramachandran plot. However, in this conformational state there is a direct hydrogen-bonding interaction between OD1 and OG in 1lga, 1ova and 2dnj, as indicated by Imberty & Perez (1995). A direct hydrogen bond is also possible in 1gal and 1cel with a small conformational change in the side chain of Ser/Thr (Table 4). Some of the direct hydrogen bonds occurring between OD1 and OG in group I and group II sequences are illustrated in Fig. 4. Group III, IV, V and VI conformations differ only in the Ψ^N dihedral angle. For these groups, the distance between OD1 and OG varies from 4 to 10 Å. In these structures, an attempt has been made to position the water molecule at hydrogen-bonding distance between OD1 and OG, keeping the main-chain conformation intact. Care is also taken not to introduce steric congestion between the position of the water molecule and the protein. It is clear that in group III such a water-mediated hydrogen bond is formed in three sequences, with a small change (if necessary) in the side-chain orientation of Ser/Thr (as given in Table 4). In 1gal, 1gly, 1hge and 1lfi, the positioning of water between OD1 and OG produces severe stereochemical clashes. In groups IV, V and VI, the location of water at the positions indicated in Table 4 facilitates the formation of water-mediated hydrogen bonding with a small change in the side chain of Ser/Thr. In the sequences numbered 2, 3 and 3 in groups IV, V and VI, respectively, such water-mediated hydrogen bonding is not possible. In most cases, this can be rectified by a small perturbation of the side chain of Asn. Of the two sequences from group VII, a water-mediated hydrogen bond is formed in one of the sequences with a small change in the side-chain orientation of Ser/Thr. Some examples of the possible water-mediated hydrogen bonds which have been described are displayed in Fig. 5. For the sequences of group VIII and IX, the conformational angles are entirely different. The formation of direct and water-mediated hydrogen bonds are unlikely in these sequences. This study implies that the direct or water-mediated hydrogen bond between the side chain of Asn and Ser/Thr may play a dominant role in the process of glycosylation in addition to the backbone conformation. This study also provides a method of

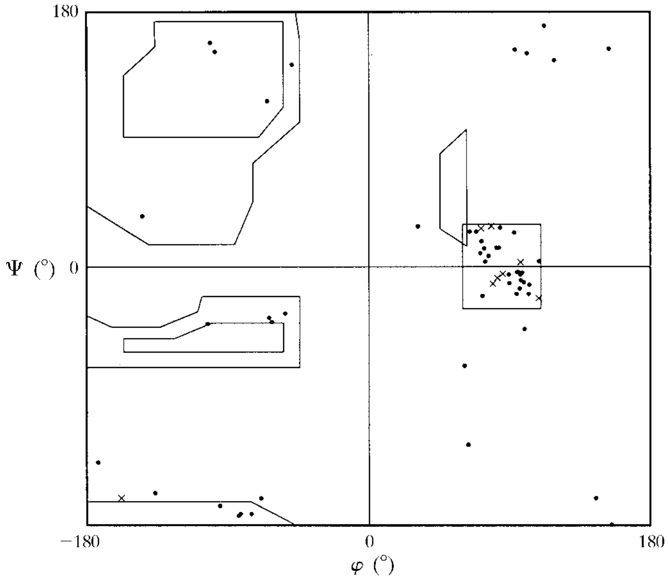


Figure 3
Ramachandran plot showing the clustered region for Gly in the glycosylatable consensus sequences. The clustered region is marked as a rectangular box. Dots and crosses represent consensus sequences from set I and confirmed glycosylated sequences, respectively.

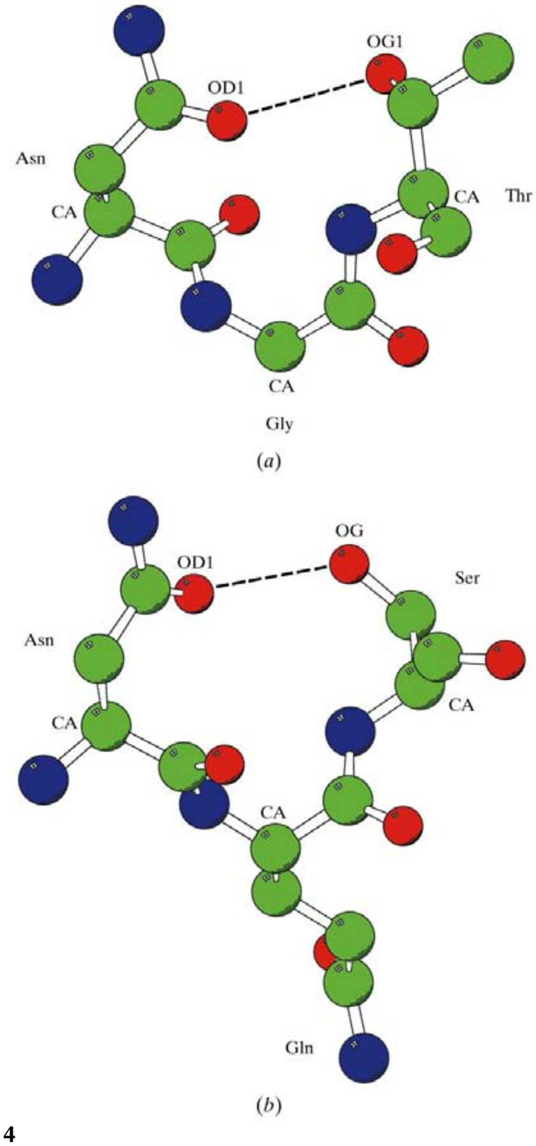


Figure 4
Illustrations of the direct hydrogen bond between OD1 and OG in the N-glycosylating sequences of (*a*) group I (1lyb) and (*b*) group II (1lga).

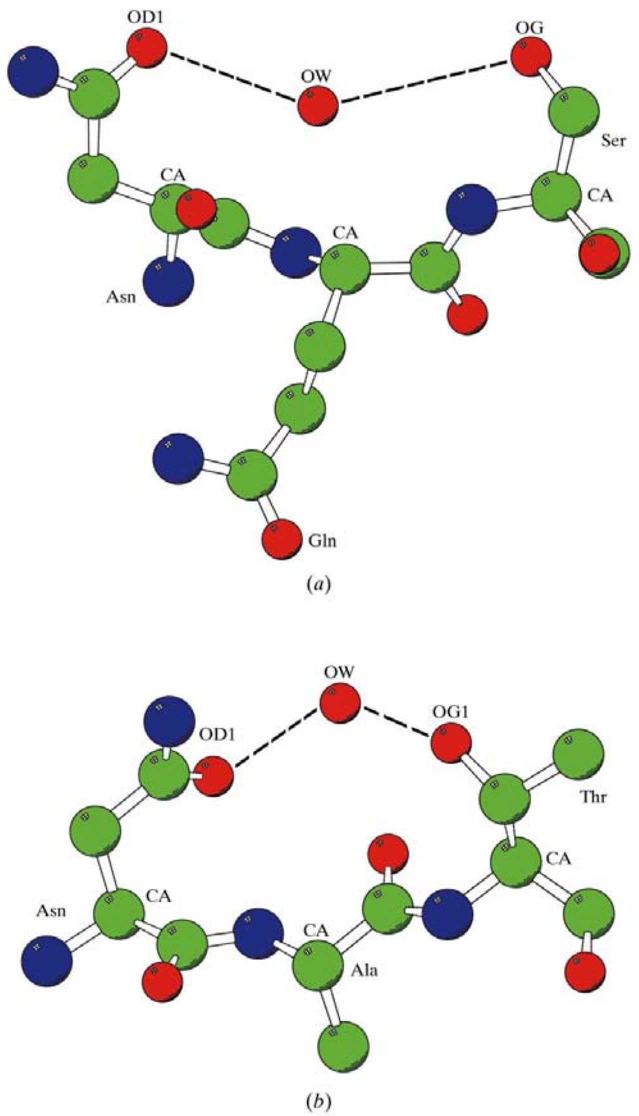


Figure 5
Figures representing the water-mediated hydrogen bond in (*a*) 2ach (group III) and (*b*) 1hge (group V). The water molecule is located at this position using the hydrogen-bond criteria; this water molecule did not cause any stereochemical clash with the protein atoms. The side-chain orientation of Ser/Thr is slightly altered in (*b*) in order to bring OG in the hydrogen-bonding position.

identifying the glycosylatable triplet sequences in globular proteins of known three-dimensional structures based on the number of spatial neighbours, conformational parameters and direct or water-mediated hydrogen bonds, which can be exploited by structural biologists and protein engineers.

The financial assistance received from DBT for this project and the use of facilities available at the Bioinformatics Centre and Interactive Graphics, Indian Institute of Science, Bangalore are gratefully acknowledged. We also wish to acknowledge Professors M. R. N. Murthy and V. S. R. Rao for fruitful discussions. THTC acknowledges the University Grants Commission for the Junior Research Fellowship.

References

Abbadi, A., Boussard, G. & Marraud, M. (1986). *Int. J. Biol. Macromol.* **8**, 252–255.

Abbadi, A., Mcharfi, M., Aubry, A., Premilat, S., Boussard, G. & Marraud, M. (1991). *J. Am. Chem. Soc.* **113**, 2729–2735.

Avanov, A. Y. (1991). *Mol. Biol. (USSR)*, **25**, 237–250.

Bause, E. (1983). *Biochem. J.* **209**, 331–334.

Bause, E. & Hettkamp, H. (1979). *FEBS Lett.* **108**, 341–344.

Bause, E. & Legler, G. (1981). *Biochem. J.* **195**, 639–644.

Bush, C. A. (1982). *Biopolymers*, **21**, 535–545.

Davis, J. T., Hirani, S., Bartlett, C. & Reid, B. R. (1994). *J. Biol. Chem.* **269**, 3331–3338.

Elbein, A. D. (1991). *Trends Biotechnol.* **9**, 346–352.

Gahmberg, C. G. & Tolvanen, M. (1996). *Trends Biochem. Sci.* **21**, 308–311.

Gavel, Y. & von Heijine, G. (1990). *Protein Eng.* **3**, 433–442.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). *Protein Sci.* **1**, 409–417.

Holst, B., Brunn, A. W., Kielland-Brandt, M. C. & Winther, J. R. (1996). *EMBO J.* **15**, 3538–3546.

Hubbard, S. C. & Ivatt, R. J. (1981). *Annu. Rev. Biochem.* **50**, 555–583.

Imberty, A. & Perez, S. (1995). *Protein Eng.* **8**, 699–709.

Imperiali, B. & Rickert, K. W. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 97–101.

Imperiali, B. & Shannon, K. L. (1991). *Biochemistry*, **30**, 4374–4380.

Imperiali, B., Shannon, K. L. & Rickert, K. W. (1992). *J. Am. Chem. Soc.* **114**, 7942–7944.

Li, Y., Luo, L., Rasool, N. & Kang, C. Y. (1993). *J. Virol.* **67**, 584–588.

Marshall, R. D. (1972). *Annu. Rev. Biochem.* **41**, 673–702.

Mononen, I. & Karjalainen, E. (1984). *Biochim. Biophys. Acta*, **788**, 364–367.

Panjikar, S. K., Biswas, M. & Saraswathi, V. (1997). *Acta Cryst.* D**53**, 627–637.

Roitsch, T. & Lehle, L. (1989). *Eur. J. Biochem.* **181**, 525–529.

Ronin, C., Bouchilloux, S., Granier, C. & Rierschoten, J. V. (1978). *FEBS Lett.* **96**, 179–182.

Veluraja, K. & Atkins, E. D. T. (1987). *Carbohydr. Polym.* **7**, 133–141.

Veluraja, K. & Mugilan, S. A. (1997). *Curr. Sci.* **72**, 572–577.