Proteomics Reveals N-Linked Glycoprotein Diversity in Caenorhabditis elegans and Suggests an Atypical Translocation Mechanism for Integral Membrane Proteins*

Hiroyuki Kaji‡§, Jun-ichi Kamiie‡, Hirotaka Kawakami‡, Kazuki Kido‡, Yoshio Yamauchi‡, Takashi Shinkawa‡, Masato Taoka‡, Nobuhiro Takahashi¶, and Toshiaki Isobe‡

Protein glycosylation is one of the most common post-translational modifications in eukaryotes and affects various aspects of protein structure and function. To facilitate studies of protein glycosylation, we paired glycosylation site-specific stable isotope tagging of lectin affinity-captured N-linked glycopeptides with mass spectrometry and determined 1,465 N-glycosylated sites on 829 proteins expressed in Caenorhabditis elegans. The analysis shows the diversity of protein glycosylation in eukaryotes in terms of glycosylation sites and oligosaccharide structures attached to polypeptide chains and suggests the substrate specificity of oligosaccharyltransferase, a single multienzyme complex in C. elegans that incorporates an oligosaccharide moiety en bloc to newly synthesized polypeptides. In addition, topological analysis of 257 N-glycosylated proteins containing a putative single transmembrane segment that were identified based on the relative positions of glycosylation sites and transmembrane segments suggests that an atypical non-cotranslational mechanism translocates large N-terminal segments from the cytosol to the endoplasmic reticular lumen in the absence of signal sequence function. Molecular & Cellular Proteomics 6:2100–2109, 2007.

Protein post-translational modifications (PTMs) such as proteolysis or addition of a chemical group to one or more amino acid residues may change the properties of a protein. A large body of evidence suggests that PTMs are critical for various cell regulatory and signaling processes (1), and thus the analysis of the status of PTMs on proteins is a major objective of proteomics research. The development of new technologies for mapping PTMs on a proteomic scale has begun to yield fruitful results (2–8). Because all PTMs accompany changes in molecular mass of a protein, MS-based analysis is often selected for large scale PTM analyses.

Among the approximately 200 different known PTMs (9), protein glycosylation is one of the most common in eukaryotes: on average there are potential targets in more than half of the genes encoded in eukaryotic genomes (10). Protein glycosylation plays a role in protein folding, subcellular localization, turnover, activity, protein-protein interactions, etc. and contributes significantly to physiology as evidenced by the growing number of human diseases with defects in glycoconjugate assembly and processing (11, 12). Thus, the analysis of protein glycosylation is important for both basic biology and clinical applications, including the discovery of protein biomarkers for diagnosis and drug discovery. Previous studies show that protein glycosylation is quite diverse because the oligosaccharide structure may vary widely between different proteins. In addition, a single protein can be glycosylated at multiple sites, and subsequent processing may differentially or partially modify an oligosaccharide attached at each site. These factors generate the observed complexity of glycoprotein structure and cause difficulties in characterizing protein glycosylation on a proteomic scale. At present, little is known about the final structure of most glycoproteins; however, the specific structure of each oligosaccharide and the rate of the modification(s) are often critical to individual glycoprotein function, and defects in these processes may cause disease (13). Thus, the mechanisms by which protein glycosylation is regulated remain a challenging problem for proteomics research.

Currently two methods allow large scale glycoprotein analysis directly from a complex biological mixture, and both methods utilize MS-based shotgun technology but differ in...
the way glycopeptides are collected. One of the methods captures glycopeptides, regardless of the glycan structure, on a solid support by chemical coupling between the cis-diol group of the glycan and hydrazide on the support, and then N-linked glycopeptides are released specifically from the support by peptide-N-glycanase (PNGase) digestion (14, 15). Another method captures a subset of glycopeptides by lectin affinity chromatography (16–18). The type of glycopeptides captured by this method depends on the specificity of the lectin used; however, comprehensive analysis of glycoproteins can be achieved by using multiple lectin columns with distinct binding specificity (e.g. non-reducing end oligosaccharides). This approach, termed isotope-coded glycosylation site-specific tagging (IGOT), includes a step to remove the glycan moiety of glycopeptides with PNGase in $^{18}$O-labeled water (16). When the enzyme releases glycans in H$_2^{18}$O, the glycosylated Asn residue (in the consensus tripeptide sequence for N-linked glycosylation, Asn-Xaa-(Ser/Thr) where Xaa is any amino acid except Pro) is converted to Asp with concomitant incorporation of $^{18}$O from water (19). This PNGase-mediated incorporation of the $^{18}$O-tag distinguishes glycosylated peptides from non-glycosylated peptides that have non-enzymatically deamidated Asp residues. The conversion of Asn to Asp via $^{18}$O incorporation in the glycosylation consensus sequence strongly indicates that the peptide was formerly N-glycosylated.

In this study, we paired IGOT with automated multidimensional liquid chromatography-MS technology and identified 1,465 N-glycosylated sites on 829 proteins expressed in Caenorhabditis elegans. We report here the diversity of protein glycosylation and the specificity of the oligosaccharyltransferase of C. elegans that incorporates an oligosaccharide moiety en bloc into nascent polypeptide chains. Based on the analysis of the relative positions of N-glycosylation sites and putative transmembrane segments of 257 potential integral membrane glycoproteins identified in this study, we also suggest that an atypical, non-cotranslational mechanism determines the topology of integral membrane glycoproteins.

**EXPERIMENTAL PROCEDURES**

*Preparation of a Galectin 6 Column*—The coding sequence of the C. elegans galectin 6 (GaL6) cDNA (provided by Dr. Hirabayashi, National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan) was inserted into the Escherichia coli expression vector pET and introduced into E. coli BL21(DE3)pLysS (20). The transformant was cultured in M9CA medium containing 0.2 mg/ml ampicillin at 37 °C, and gene expression was induced with 1 mM isopropyl 1-thio-β-o-galactopyranoside at a midlog phase of growth ($A_{600}$ = 0.6–0.8). After further cultivation for 3 h, E. coli cells were lysed by sonication at 4 °C in 50 mM sodium phosphate buffer, pH 7.5, and centrifuged at 10,000 × g for 30 min. The supernatant was then applied to an asialofetuin column (Toyopearl 650M, 2.5-cm inner diameter × 5 cm) equilibrated with 50 mM sodium phosphate buffer, pH 7.5, at a flow rate of 0.5 ml/min. After washing the column with the equilibration buffer, the adsorbed GaL6 was eluted with the same buffer containing 0.2 M lactose. The purified GaL6 (20 mg) was immobilized on TSK-GEI Tresyl-SPW (2 ml; TOSOH) according to the protocol provided by the supplier and was packed into a 4.6-mm-inner diameter × 10-cm column.

*Preparation of Tryptic Digests of Soluble and Insoluble Protein Fractions of C. elegans*—C. elegans strain N2 was cultured in liquid medium at 20 °C as described previously (16, 21). A mixed growth phase culture of the worm (5–20 g, wet weight) was lysed by sonication in 5 volumes of TBS (50 mM Tris-HCl, pH 7.5, 150 mM NaCl) containing a protease inhibitor mixture (Sigma), and the homogenate was centrifuged at 1,000 × g for 10 min at 4 °C to remove cell debris. The soluble extract was then centrifuged at 100,000 × g for 30 min at 4 °C to separate the soluble and insoluble protein fractions. Each fraction was solubilized in 7 m guanidine HCl in 0.5 M Tris-HCl, pH 8.6, containing 50 mM EDTA, and the proteins were reduced with dithiothreitol and S-carbamoylmethylated with iodoacetamide (22). The S-carbamoylmethylated proteins were dialyzed against 10 mM HEPES-NaOH, pH 7.5, and digested with N-3tosylphenylalanalyl chloromethyl ketone-treated trypsin (Pierce) at an enzyme:substrate ratio of 1:50 at 37 °C. After 18 h, an aliquot of protease inhibitor mixture (Sigma) was added to the mixture to stop digestion and to protect the lectin columns.

*Preparation of Lectin-bound Glycopeptides*—In our earlier attempts, we prepared an N-glycosylated protein fraction by lectin affinity chromatography of C. elegans crude extract and then obtained N-glycosylated peptides from a tryptic digest of the glycosylated protein fraction by a second round of lectin affinity chromatography (16). In this study, however, we modified the procedure to more efficiently identify the integral membrane glycoproteins; the crude protein extract was first digested with trypsin after S-carbamoylmethylation in 7 m guanidine HCl, and then the N-glycosylated peptides were recovered by lectin affinity chromatography. To increase the purity of glycopeptides, we incorporated an additional “hydrophilic interaction” chromatography step (23) before PNGase-mediated $^{18}$O labeling (described later).

To collect N-glycosylated peptides, the tryptic digests of soluble and insoluble protein fractions of C. elegans were subjected to affinity chromatography on three lectin columns, concanavalin A (Con A) (LA-Con A; 4.6-mm inner diameter × 15 cm; Seikagaku Corp., Tokyo, Japan), wheat germ agglutinin (WGA) (LA-WGA; 4.6-mm inner diameter × 15 cm; Seikagaku Corp.), or GaL6 (4.6-mm inner diameter × 10 cm). Approximately 50–200 mg of peptide mixture was applied to each column equilibrated with 10 mM HEPES-NaOH, pH 7.5. After washing the column with the equilibration buffer, adsorbed glycopeptides were recovered by elution with the buffer containing a cognate sugar: 0.2 M α-methyl mannospyranoside for the Con A column, 0.2 M N-acetyl-D-glucosamine (GlcNac) for the WGA column, or 0.2 M lactose for the GaL6 column. To maximize the recovery of glycosylated peptides, the flow-through fraction of the first chromatography was applied again to the same lectin column, and the chromatography was repeated as described above. The glycopeptide fractions from individual lectin columns of the first and second rounds of chromatography were combined for subsequent steps.

* Purification of Glycopeptides by Hydrophilic Interaction Chromatography*—The N-glycosylated peptide mixture recovered by lectin affinity chromatography (10–20 ml containing 200–500 μg of peptides) was added to an equal volume of ethanol (EtOH) and 4 volumes of 1-butanol (BuOH) and was applied immediately to a Sepharose CL-4B column (5-mm inner diameter × 50 mm) equilibrated with the solvent $H_2O$ : EtOH : BuOH = 1:1:4 (v/v/v). After washing the column with the same solvent, adsorbed glycopeptides were eluted with $H_2O$ : EtOH : 1:1 (v/v). The column eluent was monitored at 220 nm, and the recovered glycopeptides were quantitated fluorometrically after reaction with o-phthalaldehyde (24).

PNGase-mediated $^{18}$O Labeling of Glycopeptides—N-Glycosylated peptides were labeled specifically with $^{18}$O by IGOT as de-
C. elegans Glycoproteins and Their Membrane Topology

The glycosylation reaction takes place within the lumen of the endoplasmic reticulum (ER), N-linked glycoproteins should also have a signal sequence and/or a transmembrane segment as discussed later. We identified these two structural elements in ~25% of the 22,500 genes predicted from the genome sequence of C. elegans (Wormpep), suggesting that there are ~6,000 potential targets for N-linked glycosylation. To catalog N-glycosylated proteins expressed in C. elegans and to study details of protein glycosylation, we used ICGOT coupled with MS-based proteomics. To increase the coverage, we used three types of lectin columns with different binding specificity for the oligosaccharide attached to the polypeptide chain; thus, the columns contained immobilized Con A, WGA, and GaL6 (20), which are specific for the non-reducing end of Man, GlcNAc, and Gal, respectively. In addition, the lectin affinity chromatography was performed with tryptic peptide mixtures derived from soluble and insoluble protein fractions of C. elegans (Wormpep) crude extract (see “Experimental Procedures”). The glycopeptide mixtures were further purified by hydrophilic interaction chromatography on Sepharose CL-4B, subjected to ICGOT (i.e. N-glycanase-mediated 18O labeling), and analyzed by automated 2D nano-LC-MS/MS shotgun technology to identify 18O-labeled formerly N-glycosylated peptides. To maximize the number of identifications, the shotgun analysis was repeated three times for each peptide mixture prepared by Con A, WGA, and Gal6 affinity chromatography of the soluble/insoluble fractions. Supplemental Table 1 lists all the candidate glycosylated peptides in C. elegans identified in this study and all their MS/MS spectra are shown in Supplemental Fig. 1-1 to 1-9.

Supplemental Table 2 lists the C. elegans N-glycosylated proteins and the number of glycosylation sites identified in this study. We identified 1,204 N-glycosylated sites on 686 proteins from Con A-captured glycopeptide mixtures and likewise 474 sites on 276 proteins from WGA- and 382 sites on 330 proteins from Gal6-captured glycopeptide mixtures. After eliminating redundant identifications, we had identified 1,465 N-glycosylated sites on 829 unique proteins (Fig. 1 and...
Supplemental Table 2). The number of glycosylated sites assigned on each protein ranged from 1 to 24 with an average of 1.5. The glycoproteins we identified were quite diverse in terms of subcellular localization and function, etc., yet many (approximately 50%) were integral membrane proteins such as cell surface receptors, transporters, channels, extracellular matrix proteins, and proteases.

**Structural Heterogeneity of Oligosaccharides of the Glycoproteins**

Previous studies have shown that most glycans liberated from *C. elegans* membrane proteins contain neutral sugars and have an oligomannose-type structure and that approximately 80% of N-linked glycans in *C. elegans* have a non-reducing end mannose that is recognized by Con A (30, 31). These glycans lack sialic acid as the *C. elegans* genome has no sialyltransferase gene, implying that the glycan structure is relatively simple as compared with that of mammalian cells (32). Thus, our lectin affinity analysis of glycopeptides showed that the largest subset of N-glycopeptides was identified from the Con A-captured peptide mixtures (Fig. 1); however, the glycopeptides collected by each lectin column overlapped significantly (Supplemental Table 3). For example, 24 glycopeptides assigned for him-4 (F15G9.4) were identified using the Con A column, whereas some of these glycopeptides were also recovered from the WGA and Gal6 columns, suggesting that the him-4 product has a highly heterogeneous glycan structure. Of the 1,465 glycosylated sites we determined, 138 sites on 105 proteins were found redundantly in the peptides captured by the three lectin columns, and 317 sites on 228 proteins were found in the peptides captured by two of the lectin columns. Although a subset of those peptides should have hybrid-type glycan structures that would be recognized by multiple lectins, our study implies that most of the worm glycoproteins have complex glycans that are typical of eukaryotes. It should be noted that glycan structures may be heterogeneous not only at the protein level but also that each glycosylation site may carry a complex series of N-linked glycans if one particular peptide on a single protein is recovered by multiple lectin columns (e.g. Supplemental Table 3). Unlike these proteins, however, we also found that many proteins, such as nephrin (ZK20.6) and integrin α (F54G8.3), have relatively homogeneous glycans because multiple glycopeptides were identified only in the Con A-bound fraction, suggesting that the proteins contain a high mannose-type oligosaccharide chain(s). However, our argument, based on the binding specificity of different types of lectins, should certainly be confirmed by direct structural analysis of the oligosaccharides attached to each site of the polypeptide chain.

**Amino Acid Residues Close to the Glycosylated Site**

Although we identified ~6,000 potential targets for N-glycosylation, not all those proteins were found to be N-glycosylated, and as a matter of course, not all Asn residues in the consensus sequences were glycosylated. This suggests that local sequence elements may help determine the specificity of oligosaccharyltransferase (OST), an enzyme responsible for attachment of an oligosaccharide to the newly synthesized polypeptide. We reported previously the frequencies of the amino acid residues around the 400 glycosylated sites on Con A-bound soluble proteins in *C. elegans* (16). In the present study, the analysis was performed for 1,465 unique glycosylated sites on 829 proteins captured on the three types of lectin columns (Table I). Within the consensus tripeptide sequence for N-linked glycosylation, Thr occurs at position 3 more than twice as frequently as Ser (819 Thr versus 359 Ser). Pro does not occur at positions 2 or 4 with only one exception. We also found that Cys occurs at positions −3 to 6 at 1.5–2.5 times greater frequency than that expected from natural abundance of this residue. However, we could not detect other strong amino acid preferences around the glycosylation sites. This suggests that the nematode OST can introduce an oligosaccharide to almost any Asn-Xaa-(Ser/Thr) (Xaa ≠ Pro) sequence if the nascent polypeptide chain has a properly folded tertiary structure or meets some other criteria. The genome sequence implies that *C. elegans* may have a single OST-translocon (protein-conducting channel) complex in the ER lumen, whereas the yeast *Saccharomyces cerevisiae* has multiple OST-translocon complexes that might have different specificities for other sequences near consensus glycosylation sites (33–35). However, the factors that guide OST-mediated glycosylation remain unknown.

**Topological Analysis of Integral Membrane Glycoproteins**

**Classification of Integral Membrane Glycoproteins**—Although N-linked glycoproteins are found throughout the secretory pathway, including the cell surface, the ER, Golgi, and lysosomes, the initial attachment of N-glycans to polypeptides occurs only in the ER lumen. Thus, proteins destined for N-glycosylation contain structural elements that serve as signals that target proteins to the ER. The best characterized signal is the signal peptide, a short hydrophobic segment at the N terminus of a nascent polypeptide chain that serves as a recognition site for signal recognition particle (SRP). Upon binding the signal peptide, SRP translocates the translating ribosome to the translocon on the ER membrane. The translocon then inserts the C-terminal portion of the polypeptide chain co-translationally into the ER (36). Subsequently the signal peptide is cleaved off from the nascent polypeptide chain by the ER enzyme signal peptidase, thereby exposing the newly emerged N terminus in the lumen of the ER. Thus, the signal peptide plays a major role in the determination of the transmembrane topology of integral membrane proteins. This mechanism generates the Type I transmembrane proteins that have single transmembrane segments (Table II). Another ER targeting signal is the signal anchor, which, like
the signal peptide, contains a hydrophobic segment recognized by SRP. The internal signal anchor sequence generates Type II and Type III transmembrane proteins, which differ with regard to the membrane topology of the polypeptide chain, by inserting either the N- or C-terminal transmembrane segment into the ER (37). Unlike the signal peptide, however, the signal anchor sequence is not cleaved by signal peptidase and remains in the mature protein as a transmembrane segment.

To characterize the glycoproteins identified in this study, we searched for ER targeting signals in the amino acid sequences of C. elegans N-linked glycoproteins. We first searched for signal peptide sequences using SignalP 3.0 (neural network and hidden Markov model methods) (28) and ConPred II (module DetectSig) (29). We accepted the results of prediction only when both softwares gave identical results; otherwise we considered the results ambiguous (Table II) and did not further consider them. The number and position of transmembrane segments were predicted with ConPred II (29). The results, including several representative known glycoproteins, are summarized in Table II. Of the 829 N-linked glycoproteins identified in this study, 463 had putative signal peptides, whereas 238 did not; the remaining 128 proteins gave ambiguous results. Among the 463 signal peptide-containing proteins, 181 had a single putative transmembrane segment, 58 had multiple transmembrane segments, and 224 had no predicted transmembrane segment. Transmembrane segments were also found in 193 of 238 proteins that had no detectable signal peptide (76 with single transmembrane segments and 117 with multiple transmembrane segments). We also identified 45 “ovalbumin-like” N-linked glycoproteins with no detectable signal peptide or transmembrane sequence. Thus, our predictions suggest that the N-glycosylated proteins identified in this study included 224 secretory, 181 Type I, and 76 Type II/III proteins (Supplemental Table 4).

Analysis of Integral Membrane Glycoproteins Containing a Signal Peptide—Because the initial protein glycosylation takes place only in the ER lumen and the glycosylated segments do not cross the membrane bilayer, the structural segment around the glycosylated site must face toward the ER lumen or an equivalent topological space (e.g. the Golgi lumen). Our topological analysis of membrane glycoproteins was based on the positions of experimentally determined glycosylation sites and putative transmembrane segments on the polypeptide chains. To simplify our analysis, we focused on the 257 glycoproteins containing a single transmembrane segment. The translated polypeptide segment that follows the signal peptide is introduced into the ER lumen through the translocon. If the transmembrane segment is translated and recognized by the translocon, it exits laterally from the channel into the membrane lipid, and the C-terminal portion remains in the cytosol (36). Therefore, for single span transmembrane proteins containing a signal peptide, the N-terminal portion of the transmembrane segment resides in the ER lumen. We assigned 181 of these Type I transmem-

| Amino acid | -3 | -2 | -1 | 1 (N) | 2 (X) | 3 (S/T) | 4 | 5 | 6 |
|------------|----|----|----|------|------|--------|---|---|---|
| G          | 67 (1.07) | 65 (1.04) | 76 (1.21) | 0 | 115 (1.84) | 0 | 61 (0.97) | 62 (0.99) | 56 (0.89) |
| A          | 86 (1.17) | 75 (1.02) | 84 (1.14) | 0 | 69 (0.93) | 0 | 73 (0.99) | 70 (0.95) | 70 (0.95) |
| V          | 85 (1.17) | 70 (0.96) | 77 (1.05) | 0 | 99 (1.36) | 0 | 102 (1.40) | 81 (1.11) | 69 (0.95) |
| L          | 106 (1.04) | 118 (1.16) | 110 (1.08) | 0 | 83 (0.81) | 0 | 102 (1.00) | 88 (0.86) | 98 (0.96) |
| I          | 59 (0.81) | 84 (1.15) | 62 (0.85) | 0 | 88 (1.21) | 0 | 96 (1.32) | 99 (1.36) | 74 (1.02) |
| M          | 20 (0.65) | 32 (1.04) | 19 (0.62) | 0 | 27 (0.88) | 0 | 24 (0.78) | 18 (0.58) | 17 (0.55) |
| F          | 44 (0.76) | 80 (1.39) | 66 (1.15) | 0 | 57 (0.99) | 0 | 56 (0.97) | 59 (1.02) | 57 (0.99) |
| Y          | 45 (1.20) | 59 (1.58) | 55 (1.47) | 0 | 55 (1.47) | 0 | 37 (0.99) | 39 (1.04) | 51 (1.36) |
| W          | 10 (0.77) | 17 (1.30) | 17 (1.30) | 0 | 11 (0.84) | 0 | 21 (1.61) | 22 (1.68) | 17 (1.30) |
| S          | 65 (0.69) | 90 (0.95) | 91 (0.96) | 0 | 102 (1.08) | 359 | 94 (0.99) | 95 (1.00) | 81 (0.86) |
| T          | 76 (1.10) | 63 (0.92) | 81 (1.18) | 0 | 70 (1.02) | 819 | 79 (1.15) | 89 (1.29) | 74 (1.07) |
| C          | 41 (1.70) | 38 (1.58) | 35 (1.45) | 0 | 61 (2.53) | 0 | 36 (1.49) | 37 (1.54) | 42 (1.74) |
| N          | 51 (0.88) | 64 (1.10) | 50 (0.86) | 1178 | 59 (1.02) | 0 | 44 (0.76) | 54 (0.93) | 64 (1.10) |
| Q          | 50 (1.04) | 28 (0.58) | 54 (1.12) | 0 | 46 (0.95) | 0 | 40 (0.83) | 27 (0.56) | 36 (0.75) |
| D          | 60 (0.97) | 43 (0.69) | 58 (0.93) | 0 | 48 (0.77) | 0 | 79 (1.27) | 62 (1.00) | 78 (1.26) |
| E          | 51 (0.67) | 55 (0.72) | 54 (0.71) | 0 | 84 (1.10) | 0 | 90 (1.18) | 80 (1.05) | 112 (1.46) |
| H          | 24 (0.88) | 28 (1.02) | 30 (1.10) | 0 | 15 (0.55) | 0 | 21 (0.77) | 20 (0.73) | 23 (0.84) |
| K          | 108 (1.42) | 55 (0.72) | 71 (0.93) | 0 | 50 (0.66) | 0 | 58 (0.76) | 49 (0.64) | 56 (0.74) |
| R          | 74 (1.21) | 60 (0.98) | 46 (0.75) | 0 | 39 (0.64) | 0 | 64 (1.05) | 40 (0.65) | 52 (0.85) |
| P          | 55 (0.96) | 53 (0.92) | 42 (0.73) | 0 | 0 (0.00) | 0 | 1 (0.02) | 87 (1.51) | 51 (0.89) |

The numbers in parentheses are the relative rate estimated from the actual rate (=number of occurrences/1,178) and the expected rate from the natural abundance of each amino acid.
brane proteins, and 160 of those were actually glycosylated only at the N-terminal portion of the transmembrane segment (Fig. 2). Twelve proteins were glycosylated on the C-terminal side of the transmembrane segment, and nine were glycosylated on both sides. Thus, our ability to predict signal peptides and transmembrane segments was quite good (160/181 = 88%) for single span transmembrane proteins.

Fig. 2 indicates that the Type I transmembrane proteins generally have an extracellular or luminal N-terminal segment that is much longer than the intracellular C-terminal segment: the average length of N- and C-terminal segments was 577 and 104 residues, respectively. Typical examples having a long N-terminal segment include membrane-anchored enzymes such as phospholipase, carboxypeptidase, and UDP-glucosyltransferase, and extracellular matrix proteins such as cadherin and integrin. Conversely a number of enzymes, including tyrosine kinase, tyrosine phosphatase, and guanylate cyclase, have a long (≥100-residue) C-terminal cytosolic segment. Besides those single span transmembrane proteins, we detected many Type I transmembrane proteins with a signal peptide and multiple transmembrane segments (Supplemental Table 3).

Analysis of Integral Membrane Glycoproteins Lacking a Signal Peptide—We performed a similar analysis on the single span proteins lacking a signal peptide (Fig. 3). Of 76 proteins, 48 and 26 were assigned as Type II and Type III transmembrane proteins, respectively. Only two proteins were modified on both sides of the transmembrane segment. Fig. 3 shows the positions of the glycosylation sites and the putative transmembrane segment on the polypeptide chains of Type II and Type III proteins. In the Type II proteins, the putative transmembrane segment appeared immediately after a short N-terminal sequence. The N-terminal segments that preceded the transmembrane segment had an average length of 82 residues with 39 of 48 Type II proteins (80%) containing segments of less than 100 residues. This suggests that the transmembrane segment, or internal signal anchor, may replace the signal peptide function when the nascent polypeptide is targeted to the ER. The Type III transmembrane proteins, however, are clearly different from the Type II proteins in

| SP   | TM | type                | topology | typical proteins            |
|------|----|---------------------|----------|-----------------------------|
| 0    | 224| soluble (secreted) | Cytosol  | ER                          |
| 1    | 181| Type I              |          | R07B7.11 alpha-N-acetylgalactosaminidase |
| 2    | 58 | Type II             |          | C04F5.1 sid-1; transporter   |
| 0    | 45 | ovalbumin-like      | ZC116.3  | T05E11.6 hemoglobinase like  |
| 1    | 76 | Type II             | C38C6.2  | F26D10.9 atg-1; sugar (and other) transporter |
| 2    | 117| Type III            | B0273.4  | R13A5.1 cup-5; calcium channel |
|      | 128| ambiguous           | C18F3.2  | C57C12.5 mhp-1; ABC transporter |

The number of proteins classified into each group is indicated in parentheses. The presence of a signal peptide (SP) and the number of transmembrane segments (TM) were predicted as described in the text. Y illustrates the position of an oligosaccharide attached to the polypeptide chain. EGFR, epidermal growth factor receptor; ABC, ATP-binding cassette.
Fig. 2. Schematic representation of the topology of the putative single span transmembrane glycoproteins that contain a signal peptide sequence. The putative signal sequence is shown as an orange box at the left end of each polypeptide, and the transmembrane segment is indicated by a red box at position "1." N- (blue) and C-terminal (yellow) segments are to the left and right sides, respectively, of the transmembrane segment. The positions of potential N-glycosylation sites with a consensus sequence Asn-Xaa-(Ser/Thr) (Xaa ≠ Pro) are indicated by open diamonds (◇), and among them the experimentally determined N-glycosylated sites are indicated by closed diamonds (◇). In this figure, the Type I single span transmembrane proteins are tentatively classified into those having a small (<100-residue) or a large (>100-residue) cytosolic segment, indicated by a or b, respectively. In a number of glycoproteins (c), the glycosylated sites are located in the putative C-terminal segment or at both sides of the transmembrane segment probably due to inaccurate prediction of the transmembrane segment.
their length of the N-terminal segments. Namely Type III proteins have an average of 520 residues between the N terminus and transmembrane segment with the length ranging from 36 to 2,086 residues. Unlike the Type II proteins, 22 of 26 (85%) Type III proteins have long N-terminal segments often far exceeding 100 residues (Fig. 3). For example, UNC-5 (netrin receptor/axon guidance protein) has a 350-residue N-terminal segment that consists of an extracellular Ig-like domain (Prosite document PDOC50835 in ExPASy, www.expasy.org) and two thrombospondin domains (Prosite PDOC50092). The C01G6.8 gene product (CAM-1) has an N-terminal 450-residue segment consisting of the Ig-like, Frizzled (Prosite PDOC50038), and Kringle2 (Prosite PDOC50020) domains upstream of the putative transmembrane segment. Our prediction that UNC-5 and CAM-1 have a Type III topology is reasonable because both proteins have typical intracellular domains, ZU5 (Prosite PDOC51145) and the death (Prosite PDOC50017) domains (UNC-5) or a protein kinase domain (CAM-1), downstream of their putative transmembrane segments.

**Conclusion**

We identified a series of N-linked glycoproteins by collecting multiple subsets of N-linked glycopeptides from tryptic
C. elegans Glycoproteins and Their Membrane Topology
digests of soluble and insoluble fractions of crude extracts of C. elegans using three types of lectin columns followed by IGOT-LC-MS/MS analyses of the glycopeptides. The analysis yielded a large dataset of N-glycosylated proteins as well as the sites of glycosylation (Supplemental Table 2). Although the IGOT strategy reveals the status of glycosylation of cellular proteins, it also implies the structure of N-glycans attached to the polypeptide chain (Supplemental Table 3). The oligosaccharide chain is synthesized without a template, and the final oligosaccharide structure is determined by many factors, such as the types of glycosyltransferases and glycosidases in the cell, enzyme/substrate concentrations, and the primary and higher order structures of the substrate protein, etc. The specific structure of the oligosaccharide chain contributes to the structure/function of particular glycoproteins, and there is an expanding number of human diseases with defects in glycoconjugate assembly and processing (11, 12). This study suggests that OST, a single multienzyme complex in C. elegans, recognizes the consensus tripeptide glycosylation sequence as well as additional determinants to introduce an oligosaccharide to an Asn residue; however, it is still difficult to predict which consensus sites will undergo glycosylation and the subsequent processing that yields the final oligosaccharide structure. Clearly more extensive analysis is required to understand the glycosylation machinery of the cell.

Based on the structural characteristics of putative single span transmembrane glycoproteins identified in this study, we propose an atypical translocation mechanism for Type III transmembrane proteins in which newly synthesized polypeptide is post-translationally translocated into the ER through the translocon (Fig. 3). The mechanism underlying transmembrane protein topogenesis remains controversial (38) mainly because definitive structural data on integral membrane proteins are limited. In particular, there have been few naturally occurring Type III proteins reported (e.g. synaptotagmin II (38), microsomal cytochrome P450 (39), and yeast Golgi ecto-ATPase Ynd1p (40)), and therefore Type III transmembrane proteins are thought to have unusual topologies. It is generally accepted that the topology of Type II/III proteins is determined by the interaction between the translocon channel and the transmembrane segment of each protein essentially according to the “positive-inside” rule (38, 41). It is also believed that Type III transmembrane proteins have a relatively short N-terminal segment preceding the signal anchor or transmembrane sequence. This is due to the fact that the translocation of this segment is a post-translational event, and thus the nascent polypeptide into the translocon of the ER. Our analysis for the Type III proteins assigned here indicates that many have an N-terminal structure stabilized by disulfide bridges, such as Ig-like domain, Frizzled domain, Kringle2, FN3 (fibronectin type III), EGF3 (epidermal growth factor), Sushi, LDLRA2 (LDL-receptor class A), LDLRB (class B), and AMOP (adhesion-associated domain in MUC4 and other proteins) domains. We assume that these domains retain a relatively flexible, unfolded structure in the reducing environment of the cytosol that might be advantageous for the post-translational translocation of the polypeptide into the ER. Thus, our findings suggest that the Type III transmembrane protein architecture is widespread in eukaryotic cells. Further studies of the mechanism that determines the topogenesis of integral membrane proteins will assess the validity of these predictions.

Acknowledgment—We thank Dr. J. Hirabayashi (National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan) for the expression vector encoding worm galectin 6.

* This work was supported in part by grants for the Structural Glycomics Project from the New Energy and Industrial Technology Development Organization (NEDO) of Japan and for the Integrated Proteomics Project, Pioneer Research on Genome the Frontier from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, and by a grant-in-aid for scientific research from MEXT. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

REFERENCES

1. Mann, M., and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications, Nat. Biotechnol. 21, 255–261
2. Ballif, B. A., Villén, J., Beausoleil, S. A., Schwartz, D., and Gygi, S. P. (2004) Phosphoproteomic analysis of the developing mouse brain. Mol. Cell. Proteomics 3, 1093–1101
3. Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., Færgeman, N. J., Mann, M., and Jensen, O. N. (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol. Cell. Proteomics 4, 310–327
4. Liu, T., Qian, W.-J., Gritsenko, M. A., Camp, D. G., II, Monroe, E. M., Moore, R. J., and Smith, R. D. (2005) Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. J. Proteome Res. 4, 2070–2080
5. Liu, T., Qian, W.-J., Gritsenko, M. A., Xiao, W., Moldawer, L. L., Kaushal, A., Monroe, M. E., Vannum, S. M., Moore, R. J., Purvine, S. O., Maier, R. V., Davis, R. W., Tompkins, R. G., Camp, D., G., II., Smith, R. D., and the Inflammation and the Host Response to Injury Large Scale Collaborative Research Program (2006) High dynamic range characterization of the trauma patient plasma proteome. Mol. Cell. Proteomics 5, 1899–1913
6. Wang, L., Li, F., Sun, W., Wu, S., Wang, X., Zhang, L., Zheng, D., Wang, J., and Gao, Y. (2006) Concanavalin A-captured glycoproteins in healthy
human urine. Mol. Cell. Proteomics 5, 560–562
7. Matsumoto, M., Hatakeyama, S., Oyamada, K., Oda, Y., Nishimura, T., and Nakayama, K. I. (2005) Large-scale analysis of the human ubiquitin-related proteome. Proteomics 5, 4145–4151
8. Wykoff, D. D., and O’Shea, E. K. (2005) Identification of sumoylated proteins by systematic immunoprecipitation of the budding yeast proteome. Mol. Cell. Proteomics 4, 73–83
9. Krishna, R. G., and Wold, F. (1993) Post-translational modification of proteins. Adv. Enzymol. Relat. Areas Mol. Biol. 67, 265–298
10. 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
11. Freeze, H. H., and Aebi, M. (2005) Altered glycan structures: the molecular basis of congenital disorders of glycosylation. Curr. Opin. Struct. Biol. 15, 490–498
12. Sturiale, L., Barone, R., Fiumara, A., Perez, M., Zaffanello, M., Sorge, G., Pavone, L., Tortorelli, S., O’Brien, J. F., Jaeken, J., and Garozzo, D. (2005) Hypoglycosylation with increased fucosylation and branching of serum transferrin N-glycans in untreated galactosia. Glycobiology 15, 1268–1276
13. Moloney, D. J., Panin, V. M., Johnston, S. H., Chen, J., Shao, L., Wilson, R., Wang, Y., Stanley, P., Irvine, K. D., Haltiwanger, R. S., and Vogt, T. F. (2000). Fringe is a glycosyltransferase that modifies Notch. Nature 406, 369–375
14. Zhang, H., Li, X. J., Martin, D. B., and Aebersold, R. (2003) Identification and quantification of N-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. Nat. Biotechnol. 21, 660–666
15. Ramachandran, P., Boontheung, P., Xie, Y., Sondej, M., Wong, D. T., and Loo, J. A. (2006). Identification of N-linked glycoproteins in human saliva by glycoprotein capture and mass spectrometry. J. Proteome Res. 5, 1493–1503
16. Kaji, H., Sato, H., Yamauchi, Y., Shinkawa, T., Taoka, M., Hirabayashi, J., Kasai, K., Takahashi, N., and Isobe, T. (2003). Lactin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. Nat. Biotechnol. 21, 667–672
17. Bunkenborg, J., Pilch, B. J., Podtelejnikov, A. V., and Winsniewski, J. R. (2004). Screening for N-glycosylated proteins by liquid chromatography mass spectrometry. Proteomics 4, 454–465
18. Lewandrowski, U., Moebius, J., Walter, U., and Sickmann, A. (2006) Elucidation of N-glycosylation sites on human platelet proteins. A glycoproteomic approach. Mol. Cell. Proteomics 5, 226–233
19. Gonzalez, J., Takao, T., Hori, H., Besada, V., Rodriguez, R., Padron, G., and Shimonishi, Y. (1992) A method for determination of N-glycosylation sites in glycoproteins by collision-induced dissociation analysis in fast atom bombardment mass spectrometry: identification of the positions of carbohydrate-linked asparagine in recombinant alpha-amylase by treatment with peptide-N-glycosidase F in 18O-labeled water. Anal. Biochem. 205, 151–158
20. Hirabayashi, J., Ubukata, T., and Kasai, K. (1996) Purification and molecular characterization of a novel 16-kDa galectin from the nematode Caenorhabditis elegans. J. Biol. Chem. 271, 2497–2505
21. Mauwennyega, K. G., Kaji, H., Yamauchi, Y., Shinkawa, T., Sato, H., Taoka, M., Takahashi, N., and Isobe, T. (2003) Large-scale identification of Caenorhabditis elegans proteins by multidimensional liquid chromatography-tandem mass spectrometry. J. Proteome Res. 2, 23–35
22. Taoka, M., Yamauchi, Y., Shinkawa, T., Kaji, H., Motohashi, W., Nakayama, H., Takahashi, N., and Isobe, T. (2004) Only a small subset of the horizontally transferred chromosomal genes in Escherichia coli are translated into proteins. Mol. Cell. Proteomics 3, 780–787
23. Wada, Y., Tajiri, M., and Yoshida, S. (2004) Hydrophilic affinity isolation and MALDI multiple-stage tandem mass spectrometry of glycopeptides for glycoproteomics. Anal. Chem. 76, 6560–6565
24. Ishida, Y., Fujita, T., and Asai, K. (1981) New detection and separation method for amino acids by high-performance liquid chromatography. J. Chromatogr. 204, 143–148
25. Isobe, T., Yamauchi, Y., Taoka, M., and Takahashi, N. (2003) Automated two-dimensional liquid chromatography/tandem mass spectrometry for large-scale protein analysis, in Proteins and Proteomics (Simpson, R. E., ed.) pp. 869–876, Cold Spring Harbor Press, Cold Spring Harbor, NY
26. Natsume, T., Yamauchi, Y., Nakayama, H., Shinkawa, T., Yanagida, M., Takahashi, N., and Isobe, T. (2002) A direct nanoflow liquid chromatography-tandem mass spectrometry system for interaction proteomics. Anal. Chem. 74, 4725–4733
27. Shinkawa, T., Taoka, M., Yamauchi, Y., Ichimura, T., Kaji, H., Takahashi, N., and Isobe, T. (2005) STEM: a software tool for large-scale proteomic data analyses. J. Proteome Res. 4, 1826–1831
28. Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340, 783–795
29. Arnaout, M., Mitsuha, H., Ikeda, M., Xia, J.-X., Kikuchi, T., Satake, M., and Shimizu, T. (2004) ConsPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. Nucleic Acids Res. 32, W390–W393
30. Natsuka, S., Adachi, J., Wakuguchi, M., Nakakita, S., Hase, S., Ichikawa, A., and Ikura, K. (2002) Structural analysis of N-linked glycans in Caenorhabditis elegans. J. Biochem. 131, 807–813
31. Hirabayashi, J., Hayama, K., Kaji, H., Isobe, T., and Kasai, K. (2002) Affinity capturing and gene assignment of soluble glycoproteins produced by the nematode Caenorhabditis elegans. J. Biochem. 132, 103–114
32. Natsuka, S. (2005) Comparative biochemical view of N-glycans. Trends Glycosci. Glycotechnol. 17, 229–236
33. Yan, A., and Lennarz, W. J. (2005). Two oligosaccharide transferase complexes exist in yeast and associate with two different toxins. Glycobiology 15, 1407–1415
34. Sping, U., Bodmer, D., Wacker, M., Burda, P., and Aebersold, R. (2005) The 3.4 kDa Ost4 protein is required for the assembly of two distinct oligosaccharide transferase complexes in yeast. Glycobiology 15, 1396–1406
35. Kelleher, D. J., Karagol, D., Mandon, E. C., and Gilmore, R. (2003) Oligosaccharide transferase isoforms that contain different catalytic S1T3 subunits have distinct enzymatic properties. Mol. Cell 12, 101–111
36. Walter, P., and Johnson, A. E. (1994). Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. Annu. Rev. Cell Biol. 10, 87–119
37. Goder, V., and Spiess, M. (2001) Topogenesis of membrane proteins: determinants and dynamics. FEBs Lett. 504, 87–93
38. Kida, Y., Mihara, K., and Sakaguchi, M. (2005) Translocation of a long amino-terminal domain through ER membrane by following signal-anchor sequence. EMBO J. 24, 3202–3213
39. Sakaguchi, M., Mihara, K., and Sato, R. (1987) A short amino-terminal segment of microsomal cytochrome P-450 functions both as an insertion signal and as a stop-transfer sequence. EMBO J. 6, 2425–2431
40. Zhong, X., Malhotra, R., and Guidotti, G. (2005) A eukaryotic carboxyl-terminal signal sequence translocating large hydrophilic domains across membranes. FEBs Lett. 579, 5643–5650
41. Kida, Y., Sakaguchi, M., Fukuda, M., Mikoshiba, K., and Mihara, K. (2000) Membrane topogenesis of a type I signal-anchor protein, mouse synaptotagmin II, on the endoplasmic reticulum. J. Cell Biol. 150, 719–730