J. Biochem. 132, 103–114 (2002)

Affinity Capturing and Gene Assignment of Soluble Glycoproteins Produced by the Nematode *Caenorhabditis elegans*[1]

Jun Hirabayashi,*[,2] Ko Hayama,* Hiroyuki Kaji,† Toshiaki Isobe,† and Ken-ichi Kasai*

Department of Biological Chemistry, Faculty of Pharmaceutical Sciences, Teikyo University, Sagamiko, Kanagawa 199-0195; and †Department of Chemistry, Graduate School of Science, Tokyo Metropolitan University, Hachioji, Tokyo 192-0397

Received March 8, 2002; accepted May 7, 2002

Protein glycosylation is a central issue for post-genomic (proteomic) sciences. We have taken a systematic approach for analyzing soluble glycoproteins produced in the nematode *Caenorhabditis elegans*. The approach aims at assigning (i) genes that encode glycoproteins, (ii) sites where glycosylation occurs, and (iii) types of attached glycan structures. A soluble extract of *C. elegans*, as a starting material, was applied first to a concanavalin A (ConA) column (specific for high-mannose type *N*-glycans), and then the flow-through fraction was applied to a galectin LEC-6 (GaL6) column (specific for complex-type *N*-glycans). The adsorbed glycoproteins were digested with lysylendopeptidase, and the resultant glycopeptides were selectively recaptured with the same lectin columns. The glycopeptides were separated by reversed-phase chromatography and then subjected to sequence determination. As a result, 44 and 23 glycopeptides captured by the ConA and GaL6 columns, respectively, were successfully analyzed and assigned to 32 and 16 corresponding genes, respectively. For these glycopeptides, 49 *N*-glycosylation sites were experimentally confirmed, whereas 21 sites remained as potential sites. Of the identified genes, about 80% had apparent homologues in other species, as represented by typical secreted proteins. However, the two sets of genes assigned for the ConA and GaL6-recognized glycopeptides showed only 1 overlap with each other. Proof of the practical applicability of the glyco-catch method to a model organism, *C. elegans*, directs us to explore more complex multicellular organisms

Key words: *C. elegans*, concanavalin A, galectin, glycome, protein *N*-glycosylation.

Glycosylation endows a protein with reinforced properties not inherent to the naked protein (*1*). Because a majority of glycans are expressed on the cell surface or in the extracellular space, they are destined to play important roles in various recognition phenomena, such as in cell differentiation (*2–6*), immunity (*7*), morphogenesis (*8*), apoptosis (*9*), carcinogenesis (*10, 11*), *etc.* Ever since completion of genome/cDNA sequencing projects for some multicellular organisms, *e.g.*, *Caenorhabditis elegans* (*12*), *Drosophila melanogaster* (*13*), *Mus musculus* (*14*), and *Homo sapiens* (*15*), keen attention has been paid to protein glycosylation as a critical and complicated issue regarding post-translational modifications in eukaryotes. From a practical viewpoint, most of the genes encoding either secreted or membrane-anchored proteins are likely to be glycosylated, because they have consensus "sequeon(s)" for *N*-glycosylation sites (*i.e.*, Asn-X-Ser/Thr; X is any amino acid except Pro; *16*). In most cases, however, such information is based on computer prediction (*i.e.*, *in silico* annotation), and thus has not been confirmed by experiments. Since the process of protein glycosylation is entrusted to the eventual work of individual glycosyltransferases existing in the ER and Golgi, the final products (*i.e.*, glycoproteins) are inevitably diversified in terms of attached saccharide chains. Such heterogeneous nature of glycans is believed to have some important implications for life systems, but we are still far from arriving at a satisfactory concept. Apparently, this is due to the lack of basic knowledge on protein glycosylation from a holistic viewpoint. With the concept of "glycomics," where "glycome" is defined as a whole set of glycans produced in a single organism (*17*), our glycome projects are focused on the following four points; (i) which genes encode glycoproteins (genome information), (ii) which of their potential glycosylation sites are actually glycosylated (site information), (iii) what are individual glycan structures (structural information), and (iv) what are the biological

[1]This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Area "Genome Science" (No. 13202058 to J.H.), Grants-in-Aid for Scientific Research (No. 12680617 to J.H. and 11771453 to K.K.), and Grants for Integrated Proteomics System Project, Pioneer Research on Genome the Frontier (to T.I) from the Ministry of Education, Culture, Sports, Science & Technology of Japan, and by the Mizutani Foundation for Glycoscience.

[2]To whom correspondence should be addressed. Tel.: +81-426-85-3741, Fax: +81-426-85-3742, E-mail: j-hira@pharm.teikyo-u.ac.jp

Abbreviations: ConA, concanavalin A; EPBS, PBS supplemented with 2 mM EDTA; GaL6, *Caenorhabditis elegans* galectin LEC-6; LacNAc, *N*-acetyllactosamine; MALDI-TOF MS, matrix-assisted laser-desorption ionization time-of-flight, mass spectrometry; MEPBS, PBS supplemented with 4 mM β-mercaptoethanol and 2 mM EDTA; PA, pyridylaminated; PBS, phosphate-buffered saline (20 mM Na phosphate [pH 7.2], 150 mM NaCl); PTH, phenylthiohydantoin; SDS-PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis; TBS, Tris-buffered saline (50 mM Tris-HCl [pH 7.5], 150 mM NaCl).

effects of these glycans (functional information).

To substantiate the project, we adopted our recently devised "glyco-catch" method (*18, 19*). This enables one to assign glycoprotein genes in a systematic manner through the combination of conventional lectin-affinity chromatography and an advanced *in silico* database search. Briefly, groups of glycoproteins are purified from a crude extract on lectin-agarose columns (step 1). Since the adsorbed glycoproteins often contain non-glycoproteins (*e.g.*, through protein–protein interaction) or non-target glycoproteins (*e.g.*, through incorporation of cross-linking endogenous lectins), they are extensively digested with lysylendopeptidase (step 2). The resultant glycopeptides are reapplied to the same lectin columns as those used in step 1 (step 3). The recaptured glycopeptides are separated by reversed-phase chromatography (step 4), and then each fraction is subjected to automated sequencer analysis (step 5). Finally, nucleotide sequences satisfying the result obtained in step 5 are searched for in genome/proteome databases (step 6). For unambiguous assignment of genes, we need not know the whole sequence of the peptide. In practice, the acquisition of sequence information on 5–8 positions (not necessarily from the N-terminus) is adequate for assigning a target gene in genome/proteome databases. In principle, it is possible to begin the glyco-catch procedure with protease digestion (step 2) without purification of a group of glycoproteins (step 1). From a practical viewpoint, however, it is strongly recommended to include this step for the following reasons: firstly, if step 1 is omitted, a large amount of protease (*e.g.*, >10 mg) should be used for complete digestion of proteins included in the crude extract. Moreover, various lines of important information on target glycoproteins should be lost. These include molecular sizes determined by SDS-PAGE and reactivity with biotinylated lectins determined by lectin-blot analysis.

As the first target of this glycoproteomic approach, soluble glycoproteins secreted by the nematode *C. elegans* were chosen. Nematodes belong to the protostomes, and are thus very distant phylogenically from vertebrates, which belong to the deuterostomes. However, this worm has various merits as a multicellular model organism: firstly, its complete genome sequence is known (*12*) as well as the established cell lineage (*20*). Secondly, its genome size is relatively small (approximately 1×10^8 bp), and thus it should produce a smaller number of proteins (estimated to be <19,000) than mammals (30,000–50,000). Presently, however, only limited information is available on protein glycosylation in *C. elegans* (*21, 22*). In this study, concanavalin A (ConA) and *C. elegans* galectin LEC-6 (GaL6) were chosen to prepare distinct types of glycans; *i.e.*, high-mannose type and complex type *N*-glycans, respectively. The former is a representative plant lectin having high affinity for high-mannose type *N*-glycans (*23, 24*). On the other hand, the latter is characterized as a major endogenous galectin in *C. elegans* (*25*; for a review of galectins, see Ref. *26*). GaL6 retains all of the critical amino acids for exerting sugar-binding functions (His^{44}, Asn^{46}, Arg^{48}, Val^{59}, Asn^{61}, Trp^{69}, Glu^{71}, and Arg^{73}; the residue numbers are those of human galectin-1). Therefore, this galectin, as well as many vertebrate galectins, retains high selectivity to complex-type *N*-glycans (*19, 27, 28*). Thus, by using these two types of lectins as core tools, we can expect to acquire a global view of the glycoproteins of an organism most efficiently.

In the present study, 47 genes that encode glycoproteins recognized by either ConA or GaL6 as well as 70 *N*-glycosylation sites were assigned based on experiments. We also analyzed *N*-glycans released on hydrazinolysis from the two groups of glycoproteins, and compared their chemical structures by matrix-assisted laser-desorption ionization time-of-flight, mass spectrometry (MALDI-TOF MS).

MATERIALS AND METHODS

Reagents—ConA-agarose (5 mg/ml gel) and biotinylated ConA were products of Vector Laboratories. Anhydrous hydrazine (99%) was from Kodak. 2-Aminopyridine and all other biochemical reagents were purchased from Wako Pure Chemicals (Tokyo).

Preparation of Glycoproteins—Mixed stage worms (10 g, wet weight) were disrupted by sonication on ice with 50 ml of Tris-buffered saline (TBS; 50 mM Tris-HCl [pH 7.5], 150 mM NaCl). The homogenate was centrifuged (15,000 ×*g*, 4°C, 25 min), and the obtained supernatant (soluble extract) was applied to a ConA-agarose column (bed volume, 20 ml), which had previously been treated with 20 ml of a metal solution (0.1 mM $CaCl_2$, 0.1 mM $MnCl_2$) and then equilibrated with TBS at 4°C. After extensive washing of the column with TBS, the adsorbed glycoproteins were eluted with TBS containing 0.1 M methyl-α-D-mannoside. On the other hand, the flow-through fractions (70 ml) were pooled, supplemented with 2 mM EDTA, and then applied to a column of GaL6-agarose (bed volume, 15 ml; 5 mg/ml gel) prepared as described previously (*25*), which had previously been equilibrated with phosphate-buffered saline (PBS; 20 mM Na-phosphate [pH 7.2], 150 mM NaCl) supplemented with 4 mM β-mercaptoethanol and 2 mM EDTA (MEPBS). After extensive washing of the column, the adsorbed glycoproteins were eluted with MEPBS containing 20 mM lactose. The fraction volume was set at 20 ml throughout the experiment.

Preparation of Glycopeptides—Glycoproteins eluted from the above lectin columns (both ConA and GaL6) were precipitated with two volumes of ethanol (–20°C). After centrifugation (9,000 ×*g*, 4°C, 20 min), the precipitate was dried and protein was dissolved in 4 ml of 8 M urea containing 40 mM β-mercaptoethanol for denaturation (1 h at 37°C). Immediately after dilution of the reaction solution with 4 ml of 0.1 M $NaHCO_3$, pH 8.5, lysylendopeptidase [*Achromobacter* protease I (*29*); Wako Pure Chemicals, Tokyo], 1/100 (w/w) relative to the glycoproteins, was added, and then digestion was allowed to proceed for 48 h at 37°C. For inactivation of the protease, *p*-aminoethylbenzenesulfonyl fluoride (Wako Pure Chemicals) was then added to give a final concentration of 0.1 mM, and the incubation was continued for another 1 h at 37°C. After dilution of the digest with the same volume of equilibration buffer, TBS for the ConA fraction or EPBS for the GaL6 fraction, the digest was applied to the same type of lectin column with a reduced bed volume, *i.e.*, 10 ml (ConA-agarose) and 7.5 ml (GaL6-agarose). After extensive washing of the columns, the captured glycopeptides were eluted from the ConA column with the buffer containing 0.1 M methyl-α-D-mannoside and from the GaL6 column with that containing 20 mM lactose. The glycopeptides thus eluted (20–40 ml) were concentrated by ultrafiltration using a Centriplus YM-3 (Millipore; exclusion size, 3,000 Da), and fractionated by reversed-phase

chromatography on a trimethylsilyl (TMS)-derivatized column, Develosil UG-5 (Nomura Chemicals, Tokyo, Japan; 4.6 × 150 mm). Peptides were eluted with a linear gradient of increasing concentrations of acetonitrile in 0.1% (v/v) trifluoroacetic acid, *i.e.*, 5–15% (0–5 min) and then 15–45% (5–55 min), at a flow rate of 1.0 ml/min. The elution of peptides was monitored as to absorbance at 210 nm (A_{210}) and fluorescence (excitation and emission wavelengths, 285 and 350 nm, respectively) based on tryptophan. Fractions were taken every minute, and an aliquot of each (1/8 for ConA fractions and 1/2 for GaL6 fractions, respectively) was subjected to analysis with a Shimadzu PPSQ-21 protein sequencer.

Assignment of Genes and Prediction of Protein Types—On the basis of the peptide sequence information obtained above, genes in the *C. elegans* genome databases (GenBank/EMBL/DDBJ) were assigned essentially as described previously (*18*). For searching databases, the motif search program, SQMATCH, developed by the National Institute of Genetics, Mishima (http://ftp2.ddbj.nig.ac.jp:8080/sqmatch_ja.html), was mainly used with relatively short query sequences. On the other hand, the more common TBLAST-N (*30*) was used when searching databases with relatively long query sequences that contained multiple ambiguous sites. In any case, to establish a valid gene assignment, the following conditions must be fulfilled: (i) considering the rigorous specificity of lysylendopeptidase (*29*), the captured peptides should follow lysine except those derived from the N-terminal part of mature proteins. (ii) Each peptide should contain at least one potential *N*- or *O*-glycosylation site. (iii) The length of each peptide deduced from the genome sequence should be consistent with its retention on a reversed-phase chromatography column. Some fractions may contain more than two peptides, but it is often possible to assign multiple genes by considering the yields of phenylthiohydantoin (PTH)–amino acids and the requirements for the "hit" mentioned above.

For prediction of hydrophobic regions and classification of membrane proteins, a SOSUI program (http://www.genome.ad.jp/SOSui/index.html) was used.

Hydrazinolysis and Pyridylamination—Hydrazinolysis was performed for portions (0.1 mg) of the purified glycoprotein fractions obtained above according to an established procedure, *i.e.*, 100°C and 10 h (*31*). The liberated *N*-glycans were labeled with 2-aminopyridine to generate PA-oligosaccharides by the established procedure (*32*). An excess amount of 2-aminopyridine was removed by repeated extraction with buffered phenol (pH 7)–chloroform (1:1, v/v, 3 times) followed by chloroform (3 times), and further by gel filtration on an Asahipak GS-310 column (Asahi Chemicals, Tokyo; 7.8 × 250 mm; exclusion size determined for pullulan, 40,000 Da) equilibrated with 10 mM ammonium acetate, pH 6.0.

MALDI-TOF MS Analysis of PA-Oligosaccharides—PA-oligosaccharides derived through the above procedure were subjected to MALDI-TOF MS analysis essentially as described previously (*33*) except that dihydroxybenzoic acid was used as the matrix instead of α-cyano-4-hydroxy benzoic acid. The data were collected with a TOF mass spectrometer, PE Biosystems Voyager DE STR MALDI (Foster City, CA, USA).

RESULTS

Purification of Soluble Glycoproteins by ConA and GaL6-Affinity Chromatography—In this study, a soluble extract of worms (10 g) was first applied to a ConA-agarose column (bed volume, 20 ml), and the flow-through fractions were then applied to a GaL6-agarose column (15 ml). The proteins adsorbed on the two columns were eluted with either 0.1 M α-methyl-mannoside (ConA, shown by an arrow in Fig. 1a) or 20 mM lactose (GaL6, Fig. 2a). As a result, 23.1 and 3.9 mg, respectively, of glycoproteins were obtained. As far as examined by SDS-PAGE (Figs. 1b and 2b for the ConA and GaL6 fractions, respectively), the compositions of the observed bands for the two fractions were totally different. Therefore, it seems likely that glycoproteins specifically recognized by ConA and GaL6 were obtained. However, the possibility remains that some glycoproteins having affinity

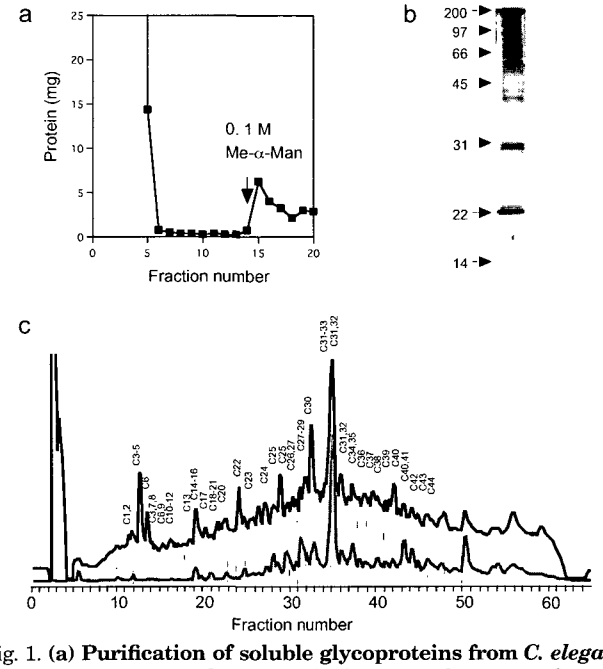

Fig. 1. **(a) Purification of soluble glycoproteins from *C. elegans* on a ConA-agarose column.** A soluble extract from 10 g of worms was applied to a column of ConA agarose (bed volume, 20 ml) equilibrated with TBS at 4°C. After extensive washing of the column, the adsorbed glycoproteins were eluted with 0.1 M methyl-α-D-mannoside (shown by an arrow). **(b) Analysis of the adsorbed glycoproteins separated by SDS-PAGE.** An aliquot of the ConA-agarose-adsorbed fraction was applied to a 14% gel, and the protein was stained with silver. **(c) Elution profile on reversed-phase chromatography of glycopeptides recaptured by ConA-affinity chromatography.** Adsorbed glycoproteins obtained in "**a**" were precipitated with EtOH, extensively digested with lysylendopeptidase, and then recaptured by using a ConA-agarose column with a reduced bed volume (10 ml). The adsorbed glycopeptides were eluted with 0.1 M methyl-α-D-mannoside (data not shown), concentrated to 4 ml by ultrafiltration (Millipore YM-3 filter), and then applied as an aliquot (2 ml) to an HPLC separation system on a TMS column (for details, see "MATERIALS AND METHODS"). An aliquot of each fraction (1 ml) was subjected to sequencer analysis, and the assigned peptides were designated as C1 to C44. For peptide sequences and relevant information, see Table I. Solid line, absorbance at 210 nm; faint line, fluorescence (excitation and emission wavelengths, 285 and 350 nm, respectively).

to GaL6, such as those having hybrid type *N*-glycans or those having both high-mannose type and complex type *N*-glycans, were adsorbed on the ConA column, because it was used first in this study. However, this possibility is less likely due to the fact that biotinylated GaL6 did not significantly stain protein bands of the ConA fraction (data not shown). The conclusion that ConA-binding glycoproteins are dominant (>80%) over GaL6-binding glycoproteins is consistent with a recent observation by Natsuka *et al.* (*34*).

Affinity Capturing of Glycopeptides and Gene Assignment—To proceed with the glyco-catch analysis, we pooled the glycoproteins described above, ethanol-precipitated them, and then digested them with lysylendopeptidase. Completion of proteolysis was confirmed by the disappearance of high-molecular-weight proteins upon SDS-PAGE on a 14% gel (data not shown), and the resultant glycopeptides were selectively recovered by using the same lectin columns. Thus recaptured glycopeptides were concentrated by ultra-filtration, and then fractionated by high-performance liquid chromatography on a trimethylsilyl reversed-phase column (Fig. 1c for the ConA-binding glycopeptides, and Fig. 2c for the GaL6-binding glycopeptides). Every fraction between 10 and 50 in each chromatography was subjected to automated sequencer analysis. Based on the acquired amino acid sequence information, corresponding nucleotide sequences were searched for in the *C. elegans* genome databases. As a result, 44 glycopeptides (C1–C44) from the ConA fraction and 23 glycopeptides (G1–G23) from the GaL6 fraction were successfully sequenced. From these data, 32 genes were assigned for the ConA fraction (Table I), whereas 16 genes were assigned for the GaL6 fraction (Table II). Notably, there was only one overlapping gene (*i.e.*, ZC64.2), which was found in both fractions (discussed more in detail later). The assigned genes are listed in Tables III and IV, respectively.

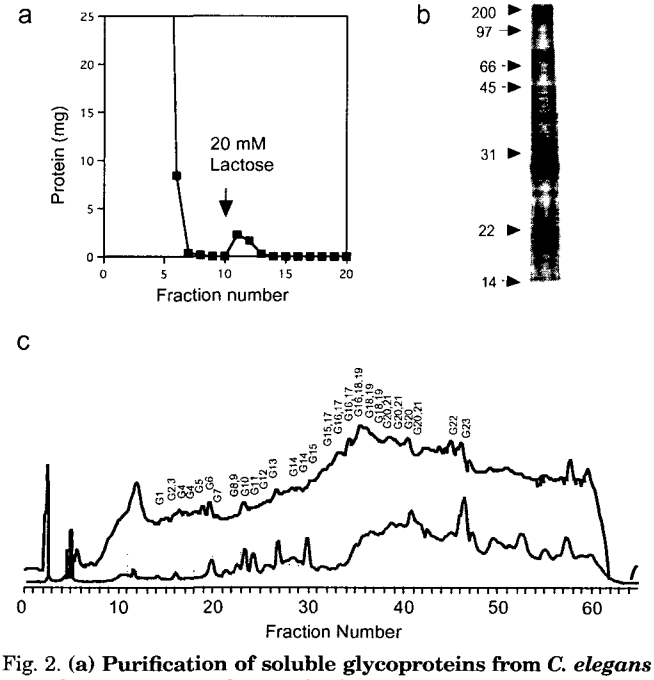

Fig. 2. **(a) Purification of soluble glycoproteins from *C. elegans* on a GaL6-agarose column.** The flow-through fraction obtained in Fig. 1a was applied to a column of GaL6 agarose (bed volume, 15 ml) equilibrated with MEPBS at 4°C. After extensive washing of the column, the adsorbed glycoproteins were eluted with 20 mM lactose (shown by an arrow). Explanations for "**b**" and "**c**" are essentially the same as in Fig. 1 except that the assigned peptides were designated as G1 to G23.

The features of the thus assigned peptides and genes are as follows: (i) As expected, these peptides were preceded by a lysine with the exceptions of C3, C19, C24, and G7. According to genome database information, all of these peptides were found to be preceded by presumed signal peptides, all of which ended with Ala. (ii) Among the analyzed 67 peptides, 60 had a single *N*-glycosylation site, and 5 (C16, G16, G17, G19, and G20) had multiple sites, whereas the remaining 2 (C39 and G18) unexpectedly lacked such a site. As a possible explanation for this discrepancy, non-glycopeptide G18 may be covalently linked to glycopeptide G19 through disulfide bond(s), because both peptides are encoded by F57F4.4, having multiple cysteine residues. More conclusively, both G18 and G19 were recovered in the same fractions, 35–38, in similar yields (see Table II) on reversed-phase chromatography (Fig. 2c). On the other hand, the lack of an *N*-glycosylation site in C39 may be attributed to either accidental contamination or miss-prediction of an open reading frame for gene F20D6.4. Yet another possibility remains, *i.e.* that C39 is a non-glycosylated peptide having significant affinity for ConA. (iii) Among the assigned 47 genes (note that ZC64.2 overlaps), 37 encoded proteins having homologues in other species and most having extracellular functions, *e.g.*, proteases (serine protease, cysteine protease, and metalloprotease), protease inhibitors (Kunitz family protease inhibitor and α2-macroglobulin), glycosidases (melibiase, β-hexosaminidase, and glucosidase), carriers for calcium ions (calsequestrin) or hormones (transthyretin), vitellogenin (*35*), extracellular type superoxide dismutase (*36*), and cell-adhesion molecules [von Willebrand factor A domain and C-type lectin (*37*)]. On the other hand, the remaining 10 had no apparent homologues in other species. (iv) According to the result of SOSUI analysis of the assigned 47 genes, 33 encoded secreted proteins as expected, having a single signal hydrophobic region in the N-terminus (known as a cleavable leader sequence). Among the remainder, 7 encoded membrane-anchored proteins, *i.e.*, 1 type-I (C44B7.5) and 3 type-II (F32B5.8, F35C5.9, and F23B2.11), and 3 with multiple membrane-anchored regions (W09D12.1, Y16B4A.2, and F57F4.4). The corresponding peptides were possibly derived through partial proteolysis during the extraction procedure, or derived from soluble proteins as a result of alternative splicing. On the other hand, the other 7 (F20D6.4, C18H7.1, T06D8.1, K07E12.1, Y49E10.18, Y4C6B.6, and T28F3.8) were predicted to be soluble, and thus cytoplasmic proteins having no hydrophobic regions. However, as far as their predicted functions are concerned, they are not likely to be cytoplasmic proteins; *i.e.*, Kunitz family protease inhibitor (F20D6.4), von Willebrand factor A domain (C18H7.1), cell-adhesion molecule (K07E12.1), lipase (Y49E10.18), and glucosyl hydrolase (Y4C6B.6) are all proteins with extracellular functions.

Representative C. elegans Glycoproteins Recognized by ConA and GaL6—Among the identified genes, C37C3.6 and F57F4.4 encode major glycoproteins in the ConA and GaL6 fractions, respectively, because both of them code for extremely large proteins (1,558 and 2,153 amino acids, respectively), contain a number of potential *N*-glycosylation

sites (Fig. 3), and are expressed abundantly (Tables III and IV). In the present analysis, 7 glycopeptides (C1, C6, C15, C23, C27, C29, and C32) from C37C3.6, and 7 (G2, G16, G17, G18, G19, G20, and G21) from F57F4.4 were captured in relatively high yields (Table I). Asn268 (C23), Asn933 (C1), and Asn1090 (C6) were not detected as a PTH derivative in the corresponding cycles, and thus were confirmed to be actually glycosylated (*i.e.*, "confirmed" glycosylation site, denoted by N on black background in Fig. 3). On the other hand, Asn445 (C29), Asn541 (C32), and Asn638 (C15) were out of the range of the sequence analysis, and therefore their glycosylation has not yet been confirmed based on experiments. However, these peptides contain only 1 potential *N*-glycosylation site; and thus they should be glyc-

TABLE I. **Summary of the assigned peptides captured by ConA-agarose.**

ID	Analyzed sequences*	Fr. No. (yield, pmol)	Cosmid
C1	NGTGCGENCL TTK.	12 (68)	C37C3.6
C2	NRTI TYDK.	12 (41)	T14F9.3
C3	RLQNVTVK.	13 (1160)	Y5F2A.1
		15 (48)	
C4	CEQHANGTIT MQAANDPK.	13 (64)	ZC412.3
C5	YLRATNRNES QLETK.	13 (80)	ZK337.1a
C6	IAACNQTQES GTVCGAGYK.	14 (188)	C37C3.6
		16 (40)	
C7	LNGSVSGLAA GK.	15 (37)	F55H2.1
C8	NYTLYK.	15 (25)	F32B5.8
C9	SICERNSTIL NARN.	16 (32)	ZK858.3
C10	SDLKPYVGNA TIK.	17 (31)	ZK337.1a
C11	NFNRTVFAED TK.	17 (28)	F40E10.3
C12	NSTDIGVYVK.	17 (38)	R07B7.11
C13	LNITNDVPAS SPADCARK.	19 (40)	T06D8.1
C14	GGLSYHGNTR NNTGIIDLSG ESK.	20 (164)	C44B7.5
C15	TRRVICAIHQ NGGLEVVDEG HCQAE*K*/ PEG*K*/ TNCTNEEK.	20 (140)	C37C3.6
C16	APENMTLPET MNTATQESAC GSL ..... NDSL ..... VPLK.	20 (64)	K07C11.3
C17	S*K*FTNETCTP HLVYSPQTAQ VK	21 (28)	C05D9.3
C18	IVNLYYNYTG DK.	22 (40)	ZK688.6
C19	GADLDVPTNI ANRTMAVK.	22 (72)	ZC64.2
C20	DGIVTGSNYT ANNGCK.	22 (112)	C25B8.3
		23 (140)	
C21	TVPADTRNLT VNVDDEDTPY VVK.	22 (19)	K07E12.1
C22	IIGCHAENVG NTIGINQNVT SGDIVYSCTK.	25 (140)	ZC412.3
C23	NGSDHFYLNG NGLIQVEK.	26 (144)	C37C3.6
C24	EITVISNLAV VGSPESHVGT PNK.	28 (112)	F54D7.4
C25	ADVPCLNDTE MLSYMNNPK.	29 (196)	F41C3.5
		30 (140)	
C26	DLGTTAFEDI SFPARAPPAF VNQTTEK.	31 (56)	C25A1.8
C27	TIAFGPHYSG CERSSFPCEL SDF ..... LG*K*/ NGTG..... LTTK.	31 (100)	C37C3.6
		32 (68)	
C28	SINVTDRFFW PPETYK.	32 (160)	T13F3.6
C29	SCETVDCEAE WFTGDWESCS STC ..... NCTV ..... PPVK.	32 (56)	C37C3.6
C30	MASVVYIESP AGVGYSYATD GNITT ..... EAVK.	33 (264)	F41C3.5
C31	TYNYWTVSSR PENNENDRVV VQL ..... NITM ..... IELK.	35 (468)	C42D8.2
		36 (120)	
		37 (56)	
C32	LLAADACPAD EQE*K*FDTERT CNL ..... NDTE ..... VTCK.	35 (108)	C37C3.6
		36 (270)	
		37 (44)	
C33	NSSSFLAPFN VQVPDEVDWR DTHLVTDVK.	35 (144)	T03E6.7
C34	RLATAYFHDE VNQTGWAFLE VDVISPK.	38 (92)	Y37D8A.2
C35	YCVGLPRGCV GTECNFAFSS ISNGTH .....V IDK.	38 (48)	C44B7.5
C36	NMIRATINVE PRQRLTVNMT IET ..... EQK.	39 (36)	K07H8.6
C37	MSSTVLNYNI TGTSSSFLIN SVE ..... SEK.	40 (64)	K07H8.6
C38	ELHDRVLFYP QGAFPLNNGS VIL ..... NDK.	41 (56)	W09D12.1
C39	SAMFVMEEP*K* DFTADHPFFF VLSK.	42 (84)	F20D6.4
C40	RARDAVSSGF EMYTVAYTA HTP ..... NE TL ..... LIDK.	43 (64)	C18H7.1
C41	ALEVGLELMA QRNVSIPTLV MVV ..... DPK.	43 (56)	C18H7.1
C42	NATLFVVNSQ DEWDAVREHF PQT ..... FEK.	44 (48)	C25B8.4
C43	TLNGSDNPAT WLQVADM VY VDN ..... DDK.	45 (64)	F22E12.1
C44	KFTVVHNYGH SGNGFTLGYG SAV..... NATISEWVK.	46 (20)	F20H11.5

*Analyzed regions are underlined. Internal lysine residues, possibly resulting from incomplete porteolysis, are denoted in italics. Asparagine residues that were confirmed to be glycosylated are shown as white N on black background, whereas those remaining as potential *N*-glycosylation sties, because they were out of the range of sequence analysis, are shown as N on gray background.

osylated from a logical viewpoint (*i.e.*, "logical" glycosylation site). Similarly, F57F4.4, Asn554, Asn840, Asn1069, and Asn1578 are considered "confirmed" *N*-glycosylation sites, whereas all of the others (*i.e.*, Asn608, Asn632, Asn656, Asn680, Asn865, Asn1192, Asn1214, Asn1789, Asn1818, and Asn1836) remain as potential *N*-glycosylation sites. In this case, no logical sites were assigned.

C37C3.6 protein belongs to the Kunitz family of protease inhibitors (PF00014 in the Pfam protein family database). It shows the highest similarity to the tissue factor pathway inhibitor precursor (*38*). On the other hand, F57F4.4 protein is characterized as a protein having EGF-like repeats (PF00008). Recent RNAi analysis of this gene showed a phenotype in which the growth of worms was suppressed, and the body became thin and sickly (*39*). As a result of a genome database search, we also found that F57F4.4 showed close similarity (96% amino acid identity) to F57F4.3 belonging to the same EGF-repeat family. Therefore, we cannot determine which gene was responsible for the captured glycopeptides. However, the latter gene (F57F4.3) is predicted to encode two additional membrane-spanning regions at the C-terminus. Notably, predicted amino acid sequences for both genes contained the GPI-anchoring consensus sequence (*40*, shown by an arrowhead

TABLE II. **Summary of the assigned peptides captured by GaL6-agarose.**

ID	Analyzed sequences*	Fr. No. (yield, pmol)	Cosmid
G1	VAAFQFANQS TAVPTYNNRR HYTRK.	15 (43)	ZK1320.2
G2	VHNNCTSPEA GVTACCCDSD ACLDPNRGK.	16 (35) 17 (49) 18 (44)	F57F4.4
G3	DVVIQAVSAN RHNYTK.	16 (10)	T28F3.8
G5	GMAFAMSDGN TTAELQK.	19 (16)	F35C5.9
G6	LVTRRNLTVT SPRQPWYYQQ GSQYVTTIAGYAK.	20 (10)	Y16B4A.2
G7	GADLDVPTNIANRTMAVK.	21 (16)	ZC64.2
G8	EGYWRSFNYN YDNAGYTSSP SENTTTC TEK.	23 (9)	R08E3.2
G9	TFVNTTIPWT RPQ SAQSVN LRES VLK.	23 (5)	F23B2.11
G10	IIADANANNSI SEPSDFFGPP K.	24 (8)	K06G5.1
G11	NFNLT VEDLQYK.	25 (13)	Y4C6B.6
G12	TFSVNCSEVG PQTDCFGFTS FDTTQK.	26 (6)	Y49E10.18
G13	RSNYSDYWIG ANDLETSGTW K.	27 (10)	W04E12.8
G14	YLEMNVTELIK.	29 (12) 30 (5)	T20D3.2
G15	YGNSSTFAVVIAPAMDLK.	31 (9)	F13H8.5
G16	NATVYYCAPV GMCRYFGLGF GGS ..... NVTA ..... NISQ..... NNSH ..... NCST .... VPGK.	32 (14) 33 (21) 34 (26) 35 (27) 36 (28) 37 (18)	F57F4.4
G17	GDCVAVNLNT TYNGVATTAS LYT ..... NSTN ..... PPRK.	32 (34) 33 (40) 34 (52)	F57F4.4
G18	DTCRPLWSDR EVTACCCNNA DNCNLK.	35 (23) 36 (24) 37 (16) 38 (8)	F57F4.4
G19	DPNVKPGPAV LPDFPTACYQ GLLVNNQTY..... NHTA..... NMSN ..... GPAK.	35 (7) 36 (10) 37 (13) 38 (13)	F57F4.4
G20	CASVNARIAN DNVTLFACVP HSL ..... NTTI ..... NGTA ..... INLK.	39 (25) 40 (17) 41 (18) 42 (20)	F57F4.4
G21	INTSMPVTNF RDYPIACFSG LYVNNMPISIAGWQACK.	39 (14) 40 (10) 41 (15) 42 (12)	F57F4.4
G22	IGNDFPLPDP IQDSTENRLI SYV.....NGTFYNGLK.	46 (15)	F40F4.6
G23	IGDDNPQIDP IQDFSSNRLI TYV.....NGTFYNGLK.	47 (13)	F28B4.3

*Analyzed regions are underlined. Internal lysine residues, possibly resulting from incomplete proteolysis, are denoted in italics. Asparagine residues that were confirmed to be glycosylated are shown as white N on black background, whereas those remaining as potential *N*-glycosylation sties, because they were out of the range of sequence analysis, are shown as N on gray background. The asparagine residue in peptide G20, which was detecetd as a PTH derivative on sequence analysis, is shown as N on rectangle.

in Fig. 3b). This type of modification has also been predicted for C05D9.3 (integrin β-subunit) and ZK337.1a (α2-macroglobulin) assigned from the ConA fraction.

MLADI-TOF MS Analysis of C. elegans N-Glycans—The above results of glyco-catch analysis directed toward soluble proteins of *C. elegans* indicated that ConA- and GaL6-captured glycopeptides had essentially no overlap. This implies that genes that encode glycoproteins having high-mannose type and complex-type *N*-glycans are largely separated by some unknown mechanism. To confirm that ConA- and GaL6-binding glycoproteins have actually distinct types of *N*-glycans, we performed hydrazinolysis of both fractions under the conditions where *N*-glycans are selectively released (*i.e.*, 100°C and 10 h; *31*). Liberated glycans were labeled with a fluorescent reagent, 2-aminopyridine (*32*), and the derived PA-oligosaccharides were directly analyzed by MALDI-TOF MS. The spectrum obtained for the PA-oligosaccharides derived from the ConA fraction included 19 peaks of significant intensity (Fig. 4a), which indicates the presence of a series of high-mannose-type *N*-glycans ranging from M3 (m/z = 1,011.51 due to $Man_3GlcNAc_2$-PA + Na^+) to M9 (m/z = 1,983.86 due to $Man_9GlcNAc_2$-PA + Na^+) structures. Notably, the spectrum also included many other peaks corresponding to M3 to M6 modified with 1 to 3 deoxyhexoses (possibly L-fucose), *i.e.*, m/z = 1,157.55 (possibly $Fuc_1Man_3GlcNAc_2$-PA + Na^+), 1,303.62 ($Fuc_2Man_3GlcNAc_2$-PA + Na^+), 1,319.64 ($Fuc_1Man_4GlcNAc_2$-PA + Na^+), 1,465.69 ($Fuc_2Man_4GlcNAc_2$-PA + Na^+), 1,481.70 ($Fuc_1Man_5GlcNAc_2$-PA + Na^+), 1,611.72 ($Fuc_3Man_4GlcNAc_2$-PA + Na^+), 1,627.78 ($Fuc_2Man_5GlcNAc_2$-PA + Na^+), 1,773.80

TABLE III. **List of the genes identified on analysis of ConA-captureed peptides and their features.**

Cosmid (Chromosome)	EST No.	ORF (aa)	Protein type	Assigned regions (peptide ID)	*N*-Glycosylation sites*	Homology to known proteins (gene name)
Y5F2A.1 (IV)	3	132	Secreted	20-27 (C3)	23	Transthyretin
K07C11.3 (V)	0	140	Secreted	68-108 (C16)	71	None
					(93)	
ZC64.2 (X)	0	164	Secreted	19-36 (C19)	30	Transthyretin
F55H2.1 (III)	0	184	Secreted	55-66 (C7)	56	Superoxide dismutase (*sod-4*)
T13F3.6 (V)	2	191	Secreted	37-52 (C28)	39	None
ZK858.3 (I)	10	225	Secreted	212-225 (C9)	217	C-type lectin
C25A1.8 (I)	113	233	Secreted	60-85 (C26)	(81)	C-type lectin
C44B7.5 (II)	0	236	Type I	37-75 (C35)	(59)	None
				187-209 (C14)	197	
ZC412.3 (V)	15	250	Secreted	154-171 (C4)	159	None
				206-235 (C22)	223	
C25B8.4 (X)	14	262	Secreted	140-177 (C42)	140	C-type lectin
F54D7.4 (I)	8	274	Secreted	22-44 (C24)	43	Two immunoglobulin domain (*zig-7*)
T03E6.7 (V)	26	337	Secreted	108-136 (C33)	108	Papain type cysteine protease
F20D6.4 (V)	0	366	Cytoplasmic	332-355 (C39)	None	Serine protease inhibitor
C05D9.3 (X)	0	372	Secreted	61-82 (C17)	65	Integrin β subunit
C25B8.3 (X)	22	379	Secreted	189-204 (C20)	196	Cysteine protease (*crp-6*)
F20H11.5 (V)	0	383	Secreted	324-379 (C44)	(371)	D-amino acidd oxidase
F40E10.3 (X)	21	417	Secreted	51-62 (C11)	53	Calsequestrin (*csq-1*)
F32B5.8 (I)	5	427	Type II	308-313 (C8)	308	Cysteine protease
R07B7.11 (V)	2	451	Secreted	317-326 (C12)	317	Melibiase
F41C3.5 (II)	2	469	Secreted	104-142 (C30)	(125)	Serine carboxypeptidase
				322-340 (C25)	328	
ZK688.6 (III)	0	507	Secreted	326-337 (C18)	331	Lysosomal Pro-X carboxypeptidase
C18H7.1 (IV)	4	534	Cytoplasmic	227-259 (C41)	239	von Willebrand factor A domain
				475-524 (C40)	(499)	
T14F9.3 (X)	2	555	Secreted	47-54 (C2)	47	β-hexosaminidase
Y37D8A.2 (I)	14	571	Secreted	51-77 (C34)	62	None
W09D12.1 (V)	4	715	Multi	615-659 (C38)	632	Collagenase
F22E12.1 (V)	8	1203	Secreted	844-877 (C43)	846	Serine carboxypeptidase
ZK337.1a (I)	2	1519	Secreted	179-811 (C10)	177	α2-macroglobulin
				1012-1026 (C5)	1019	
C37C3.6 (V)	22	1558	Secreted	268-285 (C23)	268	Kunitz type trypsin inhibitor
				395-454 (C29)	(445)	
				499-552 (C32)	(541)	
				607-632/644 (C15)	(638)	
				897-932/945 (C27)	(933)	
				933-945 (C1)	933	
				1086-1104 (C6)	1090	
C42D8.2 (X)	61	1613	Secreted	1235-1281 (C31)	1268	Vitellogenin (*vit-2*)
K07H8.6 (IV)	84	1651	Secreted	244-279 (C37)	252	Vitellogenin (*vit-6*)
				1271-1326 (C36)	1271	
T06D8.1 (II)	14	1829	Cytoplasmic	1666-1688 (C13)	1667	None
K07E12.1 (III)	0	13055	Cytoplasmic	771-793 (C21)	778	von Willebrand factor A doamin (*dig-1*)

*Potenial *N*-glycosylaton sites that were out of the range of sequenc analysis are shown in parentheses

($Fuc_3Man_5GlcNAc_2$-PA + Na^+), and 1,935.85 ($Fuc_3Man_6GlcNAc_2$-PA + Na^+). As regards such perfucosylation, the occurrence of an M3 derivative, of which the innermost GlcNAc was α1-3 and α1-6 fucosylated, and the further distal GlcNAc was α1-3 fucosylated, has previously been reported for the parasitic nematode *Haemonchus contortus* (*41*).

The spectrum also included a small number of peaks (*m/z* = 1,214.61, 1,360.68, and 1,377.66), which can be explained by incorporation of an additional *N*-acetylhexosamine (possibly GlcNAc); *i.e.*, these corresponded to $Hex_3HexNAc_3$-PA (possibly $Man_3GlcNAc_3$-PA), $Hex_3dHexHexNAc_3$-PA (possibly $Man_3Fuc_1GlcNAc_3$-PA), and $Hex_4HexNAc_3$-PA, (possibly $Man_4GlcNAc_3$-PA), respectively. They may be reaction products of either GLY-12, GLY-13, or GLY-14, which are GnT I (UDP-*N*-acetylglucosamine: α-3-D-mannoside β-1,2-*N*-acetylglucosaminyltransferase I) homologues in *C. elegans* (*42*). According to Chen *et al.*, GLY-13 is the most probable functional GnTI in *C. elegans*, and it can only utilize the physiological acceptor M5, *i.e.*, $Man_5GlcNAc_2$-Asn (*43*). Although the peak intensity was not very high, we could detect such a peak corresponding to $Hex_5HexNAc_3$-PA + Na^+ (theoretical *m/z* value, 1,538.56; denoted by an arrowhead in Fig. 4a). As regards the third peak (*m/z* = 1,377.66 corresponding to $Hex_4HexNAc_3$-PA), one of the 4 hexose residues could be galactose instead of mannose as a result of the action of some galactosyltransferase. With only this exception, no peak was observed that apparently corresponded to known complex-type *N*-glycans, such as biantennary *N*-glycans, *i.e.*, Galβ1-4GlcNAcβ1-2Manα1-3(Galβ1-4GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAc. Therefore, we can assume that *N*-glycans derived from the ConA-captured glycoproteins in *C. elegans* exclusively consisted of high-mannose type glycans.

On the other hand, the spectrum observed for the GaL6 fraction contained as many as 44 peaks of significant intensity. Among them, we could depict 12 peaks having the same m/z values as observed for the ConA fraction (*i.e.*, *m/z* = 1,157.56, 1,173.62, 1,214.56, 1,303.62, 1,319.61, 1,335.61, 1,465.67, 1,481.70, 1,611.73, 1,627.70, 1,773.79, and 1,935.84; denoted by asterisks in Fig. 4b), whereas the remaining ones were unique for the GaL6 fraction (*i.e.*, *m/z* = 954.41, 976.44, 1,037.13, 1,116.53, 1,130.55, 1,159.56, 1,278.62, 1,317.81, 1,333.64, 1,361.62, 1,418.67, 1,449.70, 1,479.70, 1,507.65, 1,625.70, 1,641.72, 1,643.74 ,1,658.74, 1,669.73, 1,771.79, 1,787.80, 1,789.77, 1,803.79, 1,815.77, 1,933.85, 1,947.88, 1,949.86, 1,975.88, 1,977.91, 2,095.87, 2,097.76, 2,137.93, and 2,139.96). At the moment, we can assign none of the latter group of peaks to any of the known complex-type *N*-glycan structures found in mammals. However, we could detect (i) the presence of multiple deoxyhexose (possibly L-fucose) residues, as in the ConA fraction, (ii) the extensive occurrence of doublet peaks with a 2-mass difference (shown in brackets in Fig. 4b), and (iii) the absence of an additional *N*-acetylglucosamine. As regards the second point, the 2-mass difference corresponds to the mass difference between a methyl ($-CH_3$) group and

TABLE IV. **List of the genes idenfied on analysis of GaL6-captureed peptides and thier features.**

Cosmid (Chromosome)	EST No.	ORF (aa)	Protein type	Assigned regions	*N*-Glycosylation sites*	Homology to known proteins (gene name)
ZC64.2 (X)	0	164	Secreted	19-36 (G7)	30	None
ZK1320.2 (II)	1	208	Secreted	71-96 (G1)	79	None
T20D3.2 (IV)	0	234	Secreted	84-94 (G14)	88	None
Y49E10.18 (III)	0	238	Cytoplasmic	6-31 (G12)	8	Lipase
R08E3.2 (X)	0	314	Secreted	184-213 (G8)	206	Proteoglycan precursor
W04E12.8 (V)	10	321	Secreted	79-99 (G13)	81	C-type lectin
F35C5.9 (II)	8	394	Type II	218-234 (G5)	227	C-type lectin
K06G5.1 (X)	30	507	Secreted	380-401 (G10)	387	None
Y4C6B.6 (IV)	0	508	Cytoplasmic	156-167 (G11)	158	Glucosyl hydrolase
T28F3.8 (IV)	4	548	Cytoplasmic	66-81 (G3)	78	None
F13H8.5 (II)	0	981	Secreted	291-308 (G15)	293	Phospholipase
F23B2.11 (IV)	23	1121	Type II	538-563 (G9)	541	Lysosomal carboxypeptidase
Y16B4A.2 (X)	21	2105	Multi	1448-1480 (G6)	1453	Lysosomal protective protein
F57F4.4 (V)	32	2153	Multi	554-709 (G16)	554 (608) (632) (656) (680)	EGF-like repeat
				832-897 (G17)	840 (865)	
				1066-1094 (G2)	1069	
				1127-1261 (G20)	(1192) (1214)	
				1577-1613 (G21)	1578	
				1738-1763 (G18)	None	
				1764-1871 (G19)	(1789) (1818,) 1836)	
F40F4.6(X)	3	2214	Secreted	1369-1421 (G22)	(1413)	EGF-like repeat
F28B4.3 (X)	13	2229	Secreted	1372-1424 (G23)	(1416)	EGF-like repeat

*Potential *N*-glycosylation sites that were out of the range of sequence analysis are shown in parentheses.

a hydroxy (-OH) group. Therefore, it is possible that some of the GaL6-recognized *N*-glycans contain methylated deoxyhexose (*e.g.*, methyl-L-Fuc, residue mass value = 160) instead of hexose (*e.g.*, galactose, residue mass value = 162). The third point completely disagrees with the previous concept that complex-type *N*-glycans contain *N*-acetyllactosamine (LacNAc) units. Since GaL6 preserves all of the essential features established for proto-type mammalian galectins, a preference for lactosamine-type disaccharides (Galβ1-3/4GlcNAc) is assured. In fact, GaL6 shows significant affinity toward a series of mammalian complex-type *N*-glycans (*19, 27, 28*). Nevertheless, the possibility remains that GaL6 recognizes disaccharide units other than LacNAc, because in many galectins the structural requirement for GlcNAc is less rigorous compared with that for galactose (*44*). For example, human galectin-1 can recognize Galβ1-4Man almost equally to Galβ1-4Glc (*45*). Moreover, it is known that *N*-acetylglucosaminidase activity is high in plants and insects, and thus the GlcNAc residue should be readily liberated even if it is transferred by GnTI (*43*). Therefore, it is quite possible that the additional GlcNAc residue observed for the three peaks for the ConA fraction (*i.e.*, m/z = 1,214.61, 1,360.68, and 1,377.66) is also liberated through the action of *N*-acetylglucosaminidase in *C. elegans*. In this case, novel complex-type *N*-glycans would be produced. There has also been a report that parasitic nematodes, such as *Schistosoma mansoni* and *Trichinella spiralis*, utilize a LacdiNAc (GalNAcβ1-4GlcNAc)

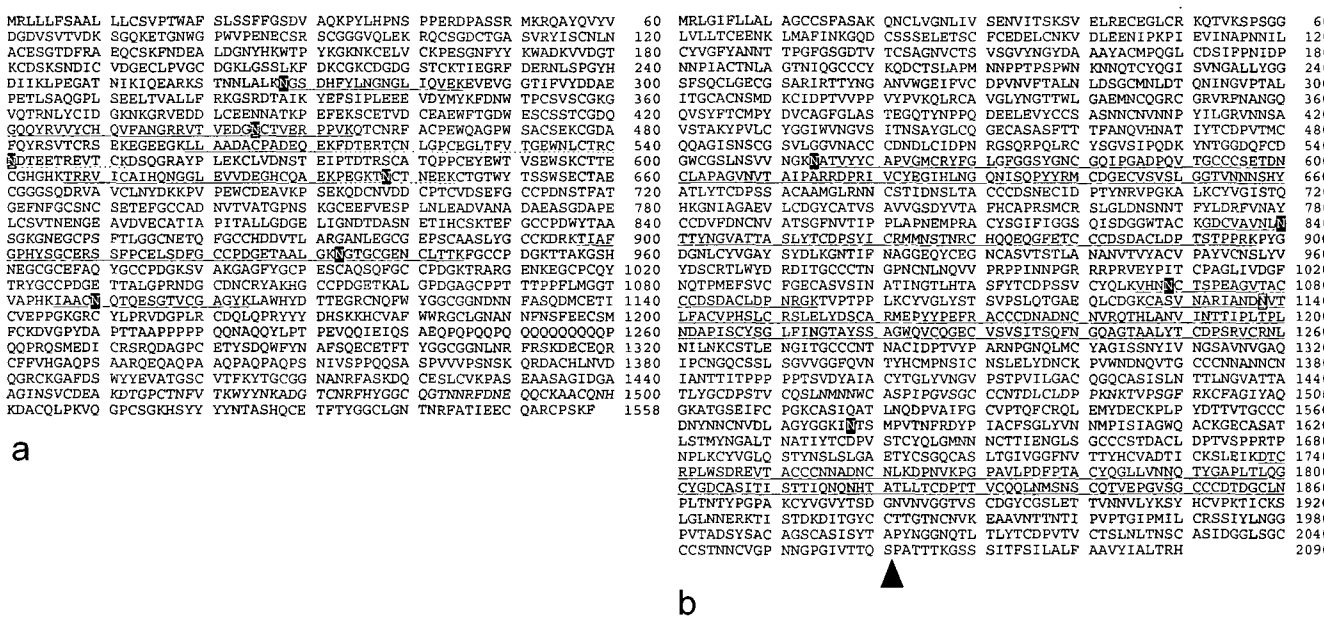

Fig. 3. **Predicted amino acid sequences for cosmids C37C3.6 (a) and F57F4.4 (b) that were identified as major glycoproteins captured by ConA and GaL6 columns, respectively.** Underline, regions corresponding to captured peptides; dashed line, extended regions beginning with internal lysine residues presumably resistant to *Achromobacter* protease I digestion; shaded regions, presumed signal sequences required for secretion; N on black background, asparagine residue, the glycosylation of which was confirmed by sequence analysis; N on gray background, asparagine residue, as a potential *N*-glycosylation site. N on rectangle, asparagine residue detected as a phenylthiohydantoin derivative on sequence analysis. The arrowhead in "**b**" indicates a predicted GPI-anchor site (*40*).

TABLE V. **Summary of the present glyco-catch analysis.**

	ConA fraction	GaL6 fraction
Total number of identified peptides	44	23
preceded by Lys	41	22
not preceded by Lys	3 (C3, C19, C24)	1 (G7)
containing internal Lys	3 (C15, C17, C39)	1 (G19)
beginning with Lys	1 (C44)	0
not ending with Lys	1 (C9)	0
carrying a single *N*-glycosylation site	42	17
carrying multiple *N*-glycosylation sites	1 (C16)	4 (G15, G16, G18, G19)
carrying no *N*-glycosylation site	1 (C39)	1 (G17)
Total number of identified genes	32	16
encoding secreted proteins	25	9
encoding type I membrane proteins	1 (C44B7.5)	0
encoding type II membrane proteins	1 (F32B5.8)	2 (F35C5.9, F23B2.11)
encoding multi-spanning membrane proteins	1 (W09D12.1)	2 (Y16B4A.2, F57F4.4)
encoding cytoplasmic soluble proteins	4 (F20D6.4, C18H7.1, T06D8.1, K07E12.1)	3 (Y49E10.18, Y4C6B.6, T28F3.8)
encoding proteins having homology to known proteins	26	11
encoding proteins having no homology to known proteins	6	5

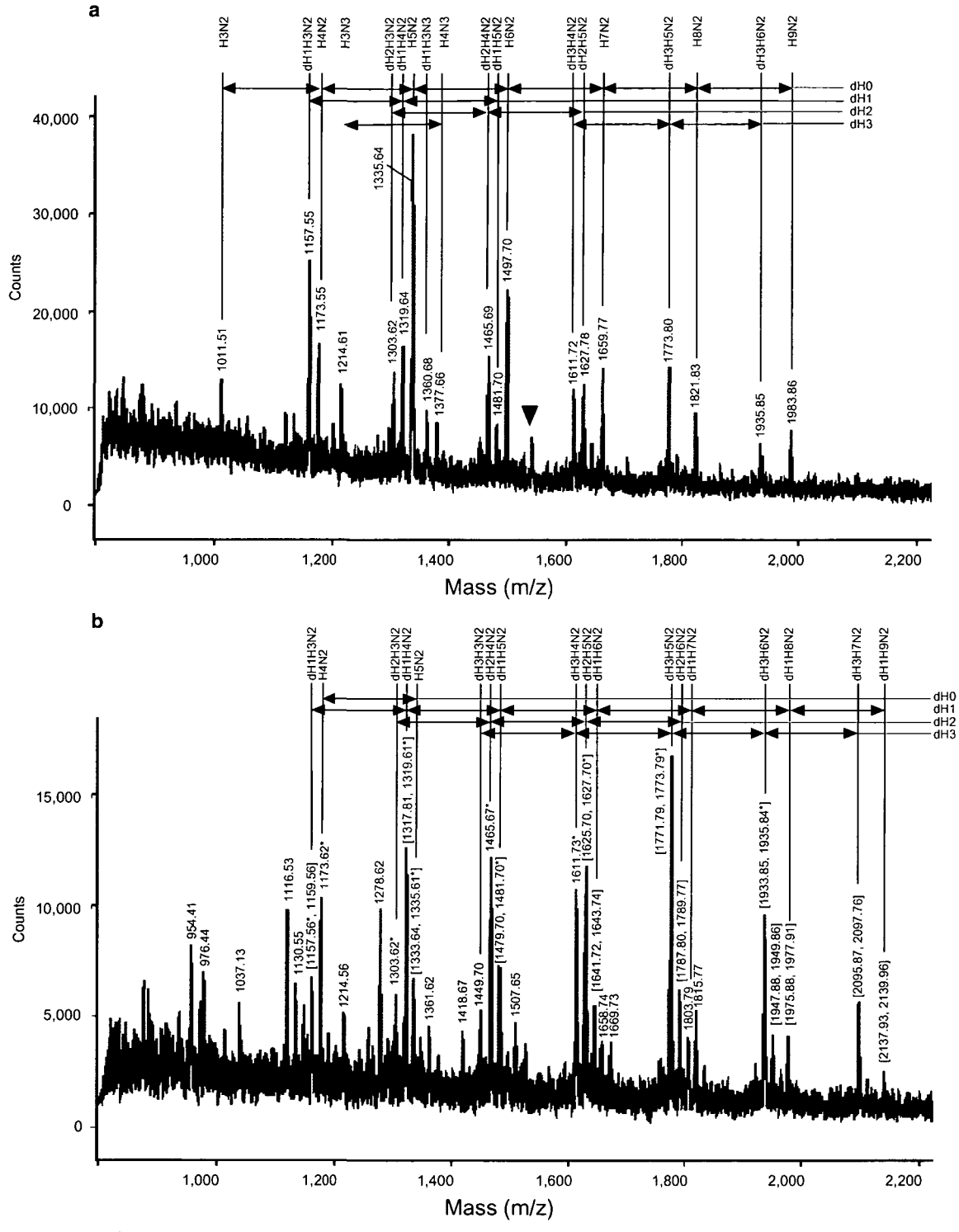

Fig. 4. **MALDI-TOF MS analysis of the total PA-oligosaccharides derived from glycoproteins adsorbed to ConA-agarose (a) and GaL6-agarose (b).** *N*-Glycans attached to the purified glycoproteins were liberated by hydrazinolysis, labeled with 2-aminopyridine, and then subjected to MALD-TOF MS analysis. Theoretically possible sugar compositions are shown above the individual mass values, where H, dH, and N represent hexose (possibly Man or Gal), deoxyhexose (possibly L-Fuc), and *N*-acetylhexosamine (possibly *N*-acetylglucosamine), respectively. Horizontal lines with arrows denote a mass difference due to to a hexose residue (*i.e.*, 162). An arrowhead in "**a**" denotes a signal corresponding to $Hex_5HexNAc_3$-PA + Na^+ (theoretical *m/z* value, 1,538.56).

unit rather than a LacNAc (Galβ1-4GlcNAc) one (*46, 47*). However, this is not likely in the case of the free-living nematode *C. elegans*, because no signal was observed that corresponded to a glycan containing this unit. Alhough further studies are necessary to define the compositions and linkage features of the complex-type *N*-glycans of *C. ele-*

gans, our conclusion is consistent with that the ConA- and GaL6-captured glycoproteins are distinct as regards both genes and glycans.

DISCUSSION

In the present study, 32 genes in the *C. elegans* genome database were assigned to 44 glycopeptides specifically captured by ConA-agarose, whereas 16 genes were assigned to 23 glycopeptides specifically captured by GaL6-agarose (for summary, see Table V). This has never been attained by means of a conventional proteomic procedure, in particular for *C. elegans*, because evidence for protein glycosylation is extremely limited in the case of *C. elegans*: for example, among 62 hits against the queries "*N*-glycosylation" or "glycosylation (unknown type)" to a recent *C. elegans* proteome database (http://www.proteome.com), only 6 cases (9.7%) are based on experimental results with some biochemical background; *i.e.*, ZK945.9 (*lov-1*), C44B12.2 (*ost-1*), T01C8.7 (*mec-4*), C42D8.2 (*vit-2*), C04F6.1 (*vit-5*), and K07H8.6 (*vit-6*). On the other hand, the remaining 56 cases are based on only prediction, and thus lack experimental confirmation. Among the former genes, C42D8.2 (*vit-2*) and K07H8.6 (*vit-6*) were previously characterized as glycoproteins recognized by ConA (*35*), consistent with the present results. On the other hand, F55H2.1 (*sod-4*), which has been predicted to be a glycoprotein only by analogy to homologous proteins in other species (*36*), proved to be actually glycosylated in this study. In this regard, the present analysis is extremely meaningful as the first demonstration of the practical validity of the proposed strategy.

Based on the present results, it is now possible to predict a substantial number of *C. elegans* genes that encode proteins that are actually glycosylated. For the sake of simplicity, the following discussion focusses on *N*-linked type glycoproteins. According to the results of a *C. elegans* data base search, the number of genes containing at least one hydrophobic region, *i.e.*, those required for secretion or membrane anchoring, is approximately 7,000. Although most of them (but not all) contain at least one potential *N*-glycosylation site, these potential sites are, from an empirical viewpoint, not always *N*-glycosylated. Thus, a rough estimate of the actual number of genes encoding *N*-linked type glycoproteins in *C. elegans* can be 5,000. On the other hand, the contribution of *O*-glycans is assumed to be much smaller; *i.e.*, approximately 10% (*16*). Among the above 7,000 genes, 2,600 are for proteins having a single hydrophobic region, *i.e.*, secreted or type II membrane proteins. If we assume that 50% of them constitute the secreted proteins, the number of genes encoding the secreted *N*-linked type glycoproteins is estimated to be 1,000: *i.e.*, 2,600 × 0.5 (the probable number of secreted proteins)–300 (the probable number of non-glycosylated proteins). The total number of genes identified in this work is 47, which corresponds to 4.7% of coverage (*i.e.*, 47/1,000 × 100%). Furthermore, if we assume that approximately 20% of the soluble glycoproteins were captured by either the ConA or GaL6 column, the ratio of successfully identified *N*-linked type glycoproteins in this work is estimated to be 24%; *i.e.*, 47/(1,000 × 0.2) × 100%. Most of the remaining 76% would be less abundant proteins or scarcely glycosylated ones. Therefore, many other minor glycoproteins should remain to be identified.

Nevertheless, the glyco-ctach method has now proved to work in principle; and thus, its application to minor components and different types of cells is promising. For example, further improvement by use of liquid chromatography/mass spectrometry (LC/MS) should enable us to identify genes for affinity-captured glycopeptides much more sensitively and with a much higher throughput. Such an experimental approach is already in progress (Kaji and Isobe, unpublished). Once glycoprotein genes are assigned, conventional glycoform analysis targeting individual glycoproteins in a more systematic manner becomes possible. Application to much more complicated organisms such as mammals is also promising. In this context, we have recently applied the glyco-catch procedure to identify genes for both soluble and membrane-anchored glycoproteins (Hirabayashi, Hashidate, and Kasai; submitted) using the recently published mouse cDNA database (*14*) and human galectin-1 (*48*) as a probe. Since the glyco-catch method depends on the binding performance of lectins, we would miss some important glycoproteins and glycopeptides having saccharide chains not fully recognized by any lectin. In this respect, we continue to pursue a new procedure for trapping every glycoprotein regardless of the structure of its saccharide chains. However, if we assume that lectins are decipherers of "glycocodes" (*49*), this defect means, on the other hand, that we can use various types of lectins, which could provide us with the "essence" of glycan information.

The important implication of the present result is that the two groups of genes identified for the ConA- and GaL6-captured fractions exhibited essentially no overlapping with each other; *i.e.*, there was only a single exception, ZC64.2 (see Tables III and IV). Apparently, this observation implies that glycoproteins having high-mannose or complex type *N*-glycans are distinguished at the gene level, as far as secreted glycoproteins are concerned. As regards membranous proteins destined to remain in the Golgi lumen, the tetrapeptide sequence "Lys-Asp-Glu-Leu" has been shown to be a Golgi retention signal (*50*). Membranous proteins having this signal remain in the lumen of the Golgi, whereas those lacking it are rapidly translocated to the plasma membrane. However, the presence of such a retention signal that directs the type of *N*-glycosylation (*i.e.*, high-mannose type or complex type) has never been discussed. To examine this point thoroughly, further studies are necessary to identify the genes that encode both secreted and membrane-anchored glycoproteins in a much more comprehensive manner. Functional analysis of individual glycosyltransferases involved in protein glycosylation (http://glycomics.scripps.edu/) is also awaited.

We wish to thank Drs. Kiyoshi Furukawa, Takeshi Sato, Shunji Natsuka, and Sumio Hase for the helpful discussions.

REFERENCES

1. Varki, A., Cummings, R., Freeze, H., Hart, G., and Marth, J., eds. (1999) *Essentials of Glycobiology*, Cold Spring Harbor Laboratoty Press, New York
2. Hakomori, S. (1992) Le^x and related structures as adhesion molecules. *Histochem. J.* **24**, 771–776
3. Metzler, M., Gertz, A., Sarkar, M., Schachter, H., Schrader, J.W., and Marth, J.D. (1994) Complex asparagine-linked oligosaccharides are required for morphogenic events during post-implan-

tation development. *EMBO J.* **13**, 2056–2065
4. Ioffe, E. and Stanley, P. (1994) Mice lacking *N*-acetylglucosaminyltransferase I activity die at mid-gestation, revealing an essential role for complex or hybrid *N*-linked carbohydrates. *Proc. Natl. Acad. Sci. USA* **91**, 728–732
5. Asano, M., Furukawa, K., Kido, M., Matsumoto, S., Umesaki, Y., Kochibe, N., and Iwakura, Y. (1997) Growth retardation and early death of β1,4-galactosyltransferase knockout mice with augmented proliferation and abnormal differentiation of epithelial cells. *EMBO J.* **16**, 1850–1857
6. Lu, Q., Hasty, P., and Shur, B.D. (1997) Targeted mutation in β1,4-galactosyltransferase leads to pituitary insufficiency and neonatal lethality. *Dev. Biol.* **181**, 257–267
7. Demetriou, M., Granovsky, M., Quaggin, S., and Dennis, J.W. (2001) Negative regulation of T-cell activation and autoimmunity by Mgat5 *N*-glycosylation. *Nature* **409**, 733–7339
8. Tsuda, M., Kamimura, K., Nakato, H., Archer, M., Staatz, W., Fox, B., Humphrey, M., Olson, S., Futch, T., Kaluza, V., Siegfried, E., Stam, L., and Selleck, S.B (1999) The cell-surface proteoglycan Dally regulates Wingless signaling in *Drosophila*. *Nature* **400**, 276–280
9. Perillo, N.L., Pace, K.E., Seilhamer, J.J., and Baum, L.G. (1995) Apoptosis of T cells mediated by galectin-1. *Nature* **378**, 736–739
10. Dennis, J.W., Laferte, S., Waghorne, C., Breitman, M.L., and Kerbel, R.S. (1987) β-1-6 Branching of Asn-linked oligosaccaharides is directly associated with metastasis. *Science* **236**, 582–585
11. Kobata, A. (1992) Structures and functions of the sugar chains of glycoproteins. *Eur. J. Biochem.* **209**, 483–501
12. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *Caenorhabditis elegans*: a platform for investigating biology. *Science* **282**, 2012–2018
13. Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., *et al.* (2001) The genome sequence of *Drosophila melanogaster*. *Science* **291**, 1304–1351
14. Kawai, J., Shinagawa, A., Shibata, V., Yoshino, M., *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690
15. McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H. *et al.* (2001) A physical map of the human genome. *Nature* **409** (2001) 934–941
16. Apweiler, R., Hermjankob, 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
17. Hirabayashi, J. and Kasai, K. (2000) Glycomics, coming of age! *Trends Glycosci. Glycotechnol.*, **12**, 1–5
18. Hirabayashi, J., Arata, Y., and Kasai, K. (2001) Glycome project: concept, strategy and preliminary application to *C. elegans*. *Proteomics* **1**, 295–303
19. Hirabayashi, J. and Kasai, K. (2002) Separation technologies for glycomics. *J. Chromatogr. B* **771**, 67–87
20. Wood, W.B. ed. (1988) *The Nematode Caenorhabditis elegans*, Cold Spring Harbor Laboratory, NY
21. Guerardel, Y., Balanzino, L., Maes, E., Leroy, Y., Coddeville, B., Oriol, R., and Strecker, G. (2001) The nematode *Caenorhabditis elegans* synthesizes unusual *O*-linked glycans: identification of glucose-substituted mucin-type *O*-glycans and short chondroitin-like oligosaccharides. *Biochem. J.* **357**, 167–182
22. Schachter, H., Chen, S., Zhang, W., Spence, A.M., Zhu, S., Callahan, J.W., Mahuran, D.J., Fan, X., Bagshaw, R.D., She, Y.-M., Rosa, J.C., and Reinhold, V.N. (2002) A functional post-translational proteomics approach to study the role of *N*-glycans in the development of *Caenorhabditis elegans*. *Brit. Biomol. J.* in press
23. Ohyama, Y., Kasai, K., Nomoto, H., and Inoue, Y. (1985) Frontal affinity chromatography of ovalbumin glycoasparagines on a concanavalin A-Sepharose column: a quantitative study of the binding specificity of the lectin. *J. Biol. Chem.* **260**, 6882–6887
24. Mega, T., Oku, H., and Hase, S. (1992) Characterization of carbohydrate-binding specificity of concanavalin A by competitive binding of pyridylamino sugar chains. *J. Biochem.* **111**, 396–400
25. 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
26. Hirabayashi, J., ed. (1997) Recent topics on galectins. *Trends Glycosci Glycotechnol.* **9**(45), 1–180
27. Hirabayashi, J., Arata, Y., and Kasai, Y. (2000) Reinforcement of frontal affinity chromatography for effective analysis of lectin-oligosaccharide interactions. *J. Chromatogr. A* **890**, 261–271
28. Hirabayashi, J., Arata, Y., Hayama, K., and Kasai, K. (2001) Galectins from the nematode *Caenorhabditis elegans* and the glycome project. *Trends Glycosci.Glycotechnol.* **13**, 533–549
29. Sakiyama, F. and Masaki, T. (1994) Lysyl endopeptidase of *Achromobacter lyticus*. *Methods Enzymol.* **244**, 126–137
30. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410
31. Takasaki, S., Mizuochi, T., and Kobata, A. (1982) *Methods Enzymol.* **83**, 263–268
32. Natsuka, S. and Hase, S. (1998) Analysis of *N*- and *O*-glycans by pyridylamination in *Methods in Molecular Biology* (Hounsell, E.F., ed.) Vol. 76, pp. 101–113, Humana Press, Totowa, NJ
33. Kaji, H., Tsuji, T., Mawuenyega, K.G., Wakamiya, A., Taoka, M., and Isobe, T. (2000) Profiling of *Caenorhabditis elegans* proteins using two-dimensional gel electrophoresis and matrix assisted laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis* **21**, 1755–1765
34. Natsuka, S., Adachi, J., Kawaguchi, M., Nakakita, S., Hase, S., Ichikawa, A., and Ikura, K. (2002) Structure analysis of *N*-linked glycans in *Caenorhabditis elegans*. *J. Biochem.* **131**, 807–813
35. Sharrock, W.J. (1983) Yolk proteins of *Caenorhabditis elegans*. *Dev. Biol.* **96**, 182–188
36. Fujii, M., Ishii, N., Joguchi, A., Yasuda, K., and Ayusawa, D. (1998) A novel superoxide dismutase gene encoding membrane-bound and extracellular isoforms by alternative splicing in *Caenorhabditis elegans*. *DNA Res.* **5**, 25–30
37. Drickamer, K. and Dodd, R.B. (1999) C-Type lectin-like domains in *Caenorhabditis elegans*: predictions from the complete genome sequence. *Glycobiology* **9**, 1357–1369
38. McVey, J.H. (1999) Tissue factor pathway. *Baillieres Best Pract. Res. Clin. Haematol.* **12**, 361–372
39. Maeda, I., Kohara, Y., Yamamoto, M., and Sugimoto, A. (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.* **11**, 171–176
40. Eisenhaber, B., Bork, P., Yuan, Y., Loffler, G., and Eisenhaber, F. (2000) Automated annotation of GPI anchor sites: case study *C. elegans*. *Trends. Biochem. Sci.* **25**, 340–341
41. Haslam, S.M., Coles, G.C., Munn, E.A., Smith, T.S., Smith, H.F., Morris, H.R., and Dell, A. (1996) *Haemonchus contortus* glycoproteins contain *N*-linked oligosaccharides with novel highly fucosylated core structures. *J. Biol. Chem.* **271**, 30561–30570
42. Chen, S., Zhou, S., Sarkar, M., Spence, A.M., and Schachter, H. (1999) Expression of three *Caenorhabditis elegans N*-acetylglucosaminyltransferase I genes during development. *J. Biol. Chem.* **274**, 288–297
43. Chen, S., Spence, A.M., and Schachter, H. (2001) Complex *N*-glycans: when, why? *Trends Glycosci. Glycotechnol.* **13**, 447–462
44. Hirabayashi, J., Hashidate, T., Arata, Y., Nishi, N., Nakamura, T., Hirashima, M., Urashima, T., Oka, T., Futai, M., Muller, W.E.G., Yagi, F., and Kasai, K. (2002) Oligosaccharide specificity of galectins: a search by frontal affinity chromatography. *Biochim. Biophys. Acta*, in press
45. Walzel, H., Hirabayashi, J., Kasai, K., Brock, J., and Neels, P. (1996) Cell calcium signalling induced by endogenous lectin carbohydrate interaction in the Jurkat T cell line. *Glycoconjugate J.* **13**, 99–106
46. Appleton, J. (2001) Novel *N*-glycans of the parasitic nematode *Trichinella spiralis*. *Trends Glycosci. Glycotechnol.* **13**, 481–492
47. Khoo, K.-H. (2001) Structural variations in Schistosomal glycans. *Trends Glycosci. Glycotechnol.* **13**, 493–506
48. Hirabayashi, J. and Kasai, K. (1991) Effect of amino acid substitution by sited-directed mutagenesis on the carbohydrate recognition and stability of human 14-kDa β-galactoside-binding lectin. *J. Biol. Chem.* **266**, 23648–23653
49. Kasai, K. and Hirabayashi, J. (1996) Galectins: a family of animal lectins that decipher glycocodes. *J. Biochem.* **119**, 1–8
50. Lewis, M.J. and Pelham, H.R. (1990) A human homologue of the yeast HDEL receptor. *Nature* **348**, 162–316