Journal of Biomolecular Techniques
13:205–218 © 2002 ABRF

REVIEWS

Glyco-Catch Method: A Lectin Affinity Technique for Glycoproteomics

Jun Hirabayashi, Tomomi Hashidate, and Ken-ichi Kasai

Department of Biological Chemistry, Faculty of Pharmaceutical Sciences, Teikyo University, Sagamiko, Kanagawa, Japan

Protein glycosylation is a critical issue of post-genome science not only because it is one of the major post-translational modifications but also because it has significant effects on protein properties and functions. The glyco-catch method was recently developed as a novel affinity technique for comprehensive analysis of glycoproteins in the context of glycomics, which is defined as research targeting the whole set of glycans produced in an organism (Hirabayashi J, Kasai K, *Trends Glycosci Glycotechnol* 2000;12:1–5). This method enables us to identify possible glycoprotein genes as well as glycosylation sites in a systematic manner by combining conventional lectin affinity chromatography and concurrent *in silico* database searching (Hirabayashi J, Kasai K, *J Chromatogr B* 2002; 771:67–87). Application of the strategy to a simple organism, *Caenorhabditis elegans,* has already proved its practical validity (Hirabayashi J, Kaji H, Isobe T, Kasai K, *J Biochem (Tokyo)* 2002;132:103–114). Accumulation of data on protein glycosylation in a variety of organisms for which entire genome information is available should thus reveal the biological meaning of glycans in complex carbohydrates from a global viewpoint, that is, under the concept of "genome-proteome-glycome." In this article, we briefly review the issues of protein glycosylation and demonstrate the usefulness of the glyco-catch method for identification of complex-type *N*-glycoproteins of mouse liver that were captured by galectin-1, which is a major galectin in mammals. Future plans for technical improvement and construction of a glycome database are also described. (J Biomol Tech 2002;13:205–218)

KEY WORDS: glycomics, galectin, glycoprotein, glyco-catch method, complex-type *N*-glycan.

ADDRESS CORRESPONDENCE AND REPRINT REQUESTS TO: Jun Hirabayashi, Research Center for Glycoscience, National Institute of Advanced Industrial Science and Technology, AIST Tsukuba Central 6, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan (email: jun-hirabayashi@aist.go.jp).

THE STUDY OF GLYCANS

Inherent Problems of Glycans

Glycosylation endows a protein with various reinforced properties that the naked protein lacks. These properties include quality control, stability, functional efficiency, destination, and cooperation with other biomolecules such as lectins.[1] In other words, it makes the role of a protein multidimensional. However, in contrast to protein phosphorylation, for which analytical strategies have already been established, approaches to protein glycosylation have many issues yet to be resolved in terms of both principle and practice. First of all, glycan structures are extremely diverse. In general, glycans found in natural glycoproteins consist of aldohexoses (e.g., glucose, mannose, galactose), their *N*-acetyl derivatives (e.g., GlcNAc and GalNAc), and in many cases deoxyhexose (e.g., L-fucose) and sialic acids represented by *N*-acetylneuraminic acid (NeuAc) (for representative component saccharides and *N*-glycan structures, see Fig. 1). The glycan moieties of proteoglycans (e.g., heparan sulfate, heparin, dermatan sulfate, chondroitin sulfate) include xylose in the so-called core region. However, complexity of glycans is not attributed to diversity of these component saccharides but rather to diversity in linkage isomers. For example, there are eight such isomers to link two aldohexoses: "α1-2/3/4/6" and "β1-2/3/4/6." Further, multiple glycoside bonds can be formed for each monosaccharide. According to Laine,[2] six monosaccharides can

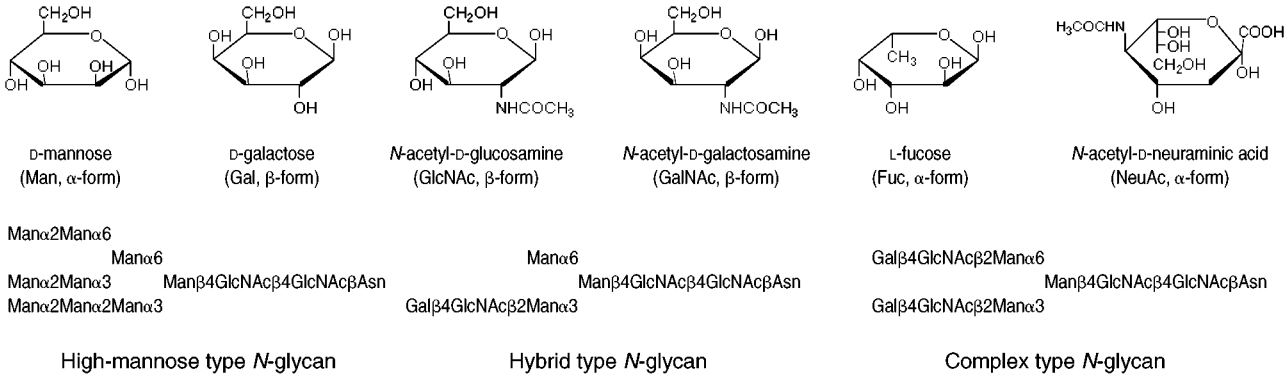

FIGURE 1

Representative component saccharides and *N*-glycan structures found extensively in animals.

make as many as 1.05×10^{12} molecules of structural complexity, which far exceeds the numbers for hexanucleotides (i.e., $4^6 = 4096$) and hexapeptides (i.e., $20^6 = 64{,}000{,}000$). Therefore, it is not surprising that there has been no practical suggestion for creating a glycan sequencer and a glycan synthesizer!

Another reason for difficulty in studying glycans is that they are not direct products of genes; rather, they are synthesized as a result of multiple steps of specific reactions to form glycosidic bonds catalyzed by glycosyltransferases. However, these reactions are not always assured to be complete due to various factors (e.g., relatively low reaction velocity and low affinity for substrates), and thus the final products become a mixture of heterogeneous forms of glycans. In addition, there often exist several transferases competing for a common acceptor saccharide. Thus, products become structurally even more heterogeneous. These facts mean that it is practically impossible to predict final glycan structures only from genome information. Such features make a clear contrast to those of proteins, the direct products of genes.

Two Ways of Approaching Glycans

As discussed above, structural diversity is a basic feature inherent to glycans, and it is often related to biological phenomena, such as development, carcinogenesis, and microbial infections.[1] Therefore, it is often necessary to investigate heterogeneous features of glycan chains of certain glycoproteins in the context of structure–function studies. Such an approach is called "glycoform analysis." In general, however, this approach requires highly expert knowledge and experience on complex glycans. On the other hand, as a result of the rapid progress in genome sequencing of an extensive number of organisms, it has become evident that there is a significant number of genes whose functions are totally unknown. If there is annotation for these function-unknown genes with respect to glycosylation (e.g., high-mannose type *N*-glycans, complex type *N*-glycans, core 1 *O*-glycans, etc.), this may provide a key to solve their functions, because certain types of glycans are targets of various recognition proteins, such as lectins. In light of this kind of approach, one may have in mind a very basic question: "How many genes in an organism encode glycoproteins?" Hence, it becomes necessary to develop a new method to register all kinds of glycoprotein genes. Such an attempt has recently been taken under the proposed concept of the "glycome," which is defined as the whole set of glycans produced in an organism.[3–5]

In this review, we focus on a new affinity technique that was developed recently, named the "glycocatch" method.[4] The method can be a core affinity-capturing technique to assign glycoprotein genes as well as glycosylation sites in a systematic manner. According to a previously established protocol of the method, (1) target glycoproteins are enriched by means of lectin-affinity chromatography (lectins are binding proteins that are specific for certain types of carbohydrates); (2) the obtained glycoproteins are extensively digested with rigorously lysine-specific lysylendopeptidase; (3) from the digest, target glycopeptides are selectively recovered by using the same lectin column as used in step 1; (4) thus-recaptured glycopeptides are separated by reversed-phase high-performance liquid chromatography (HPLC); (5) each HPLC fraction is subjected to structural analysis with a protein sequencer; and (6) on the basis of the acquired sequence information, genome databases are searched for genes that encode corresponding glycoproteins. If detailed sugar-binding specificity of the used lectin is known,[6] one can have basic information on target glycan struc-

tures (e.g., high-mannose type, complex type, etc.). Such a comprehensive approach targeting glycoproteins can be defined as "glycoproteomics."[7]

Another important issue of glycomics is to determine glycan structures. So far, various techniques have been used for this purpose: they include composition analysis, methylation analysis, glycosidase treatment, mass spectrometry (MS), and nuclear magnetic resonance (NMR). In general, however, these procedures are laborious and time-consuming, and thus not appropriate for comprehensive analysis. Nevertheless, if target glycans are derived from mammals, it is possible that they are identical to any of the known structures. Among such "glycan matching" procedures, the two-dimensional (2D) mapping system developed by Takahashi and her coworkers[8] has a basic advantage over other methods, because >400 pyridylaminated oligosaccharides, the labeling method for which was developed by Hase and his coworkers,[9] have already been mapped with coordinated retention times in both normal and reversed-phase chromatographies (http://www.gak.co.jp/ECD/Hpg_eng.htm). Recently, a combined strategy for a comprehensive purpose has been presented,[10] where three independent methods—MS, 2D mapping, and reinforced frontal affinity chromatography (FAC)[6,11,12]—are used. Because these methods are based on distinct principles (i.e., physics [MS], chemistry [2D mapping], and biochemistry [FAC]), their combination should be successful for specification (not necessarily identification to determine covalent structures based on logic) of both known and unknown glycans.

OUTLINES OF EXPERIMENTAL PROCEDURES

Glycoform Analysis

In the case of target proteins, they are usually purified first and then their structures are analyzed by either Edman degradation or MS/MS procedures. To assign corresponding genes, one can analyze functions and three-dimensional (3D) structures of the target proteins preliminarily by *in silico* searching. If they are categorized as extracellular proteins—that is, having hydrophobic signal sequences for secretion (i.e., secreted proteins) or membrane anchoring (i.e., membrane proteins)—and if they have a consensus sequence(s) for attachment of *N*-glycans (i.e., Asn-X-Ser/Thr; X is any amino acid except Pro), they are possibly glycosylated. However, because such potential *N*-glycosylation sites are not always glycosylated (the reason is not clear at the moment), it must be confirmed by experiments whether these sites are actually glycosylated or not. If so, further analysis is necessary to examine what types of glycans are attached, how large the molecular diversity is, and what biological functions they have. Overall procedures of glycoform analysis can be summarized as follows:

> Purification of target glycoproteins → Structural analysis (Edman or MS/MS) → Gene identification → Prediction of 3D structure, biological function, etc. → Liberation of glycan chains → Glycoform analysis → Structure–function analysis

Glycoproteomics

In other cases, groups of glycoproteins are analyzed systematically with no particular targets. For this purpose, however, there is no specific, single procedure for collecting all kinds of glycoproteins. Therefore, in our proposed strategy of "glyco-catch," certain types of glycoproteins are group-purified by means of lectin affinity chromatography (see Fig. 2). The overall procedures may be summarized as follows:

> Extraction of (glyco)proteins → Group purification of glycoproteins on lectin columns → Glyco-catch procedure → Assignment of glycoprotein genes and glycosylation sites

Before proceeding with either glycoform analysis or glycoproteomics, investigators are strongly urged to determine in advance whether target proteins are glycosylated or not by either periodic acid–Schiff staining or lectin-blot analysis. For the former, a convenient glycoprotein detection kit is available (e.g., BioRad Glycoprotein Immunoblot kit, Hercules, CA). The latter, lectin-blot analysis, is an application of the Western blot technique wherein biotinylated lectins are used as probes to detect specific types of glycans. Various plant lectins can be used for this purpose, for example, concanavalin A (ConA), which is specific for high-mannose type *N*-glycans; *Ricinus communis* agglutinin I (RCA-I), which is specific for *N*-acetyllactosamine-containing saccharides (i.e., represented by complex type *N*-glycans); and peanut agglutinin (PNA), which is specific for Galβ1–3GalNAc (T-antigen), a basic structure designated "core 1" and commonly included in *O*-glycans. In each experiment, one should prepare a pair of nitrocellulose blots, one of which is reacted with a probe lectin and the other used for a control treated similarly except for the addition of an appropriate competitive sugar (e.g., methyl-α-D-mannoside in the case of ConA, and lactose in the case of RCA-I). When evaluating results of lectin-blot analysis, however, one should consider carefully the detailed sugar-binding specificity of the lectin used.[10]

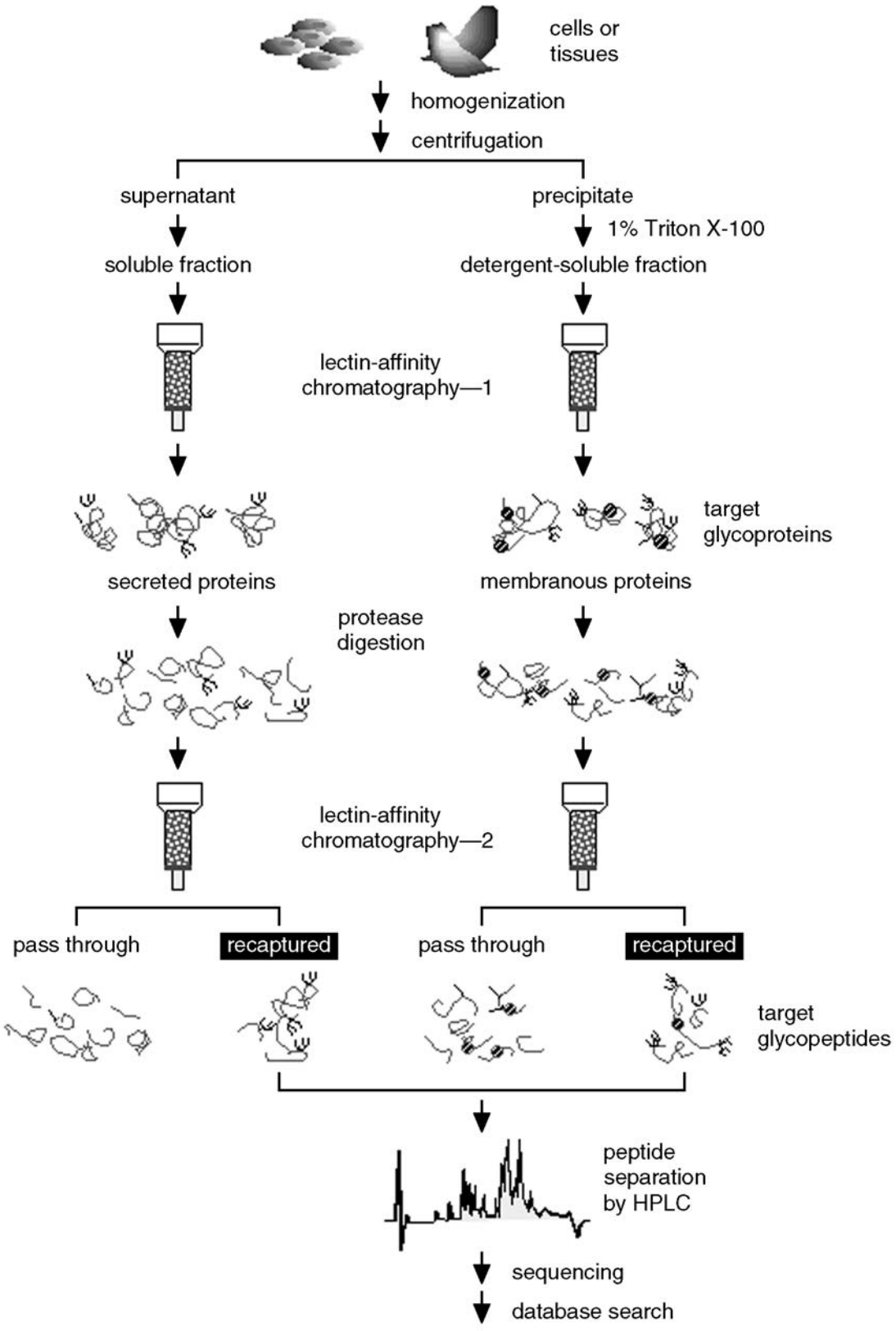

FIGURE 2

Overall procedures of the glyco-catch method. Schemes for analysis of both secreted proteins (soluble fraction; left) and membrane-anchored proteins (detergent-soluble fraction; right) are shown.

GLYCO-CATCH METHOD

Principles and Features

Procedures for the glyco-catch method are largely divided into two parts (Fig. 2). In the first part, a group of glycoproteins having certain glycan structures is purified by lectin-affinity chromatography, and the thus-purified glycoproteins are then extensively digested with an appropriate protease. The resultant glycopeptides are recaptured on the same lectin column. There are three reasons why it is necessary to have glycopeptides but not glycoproteins: (1) Selectivity for target glycans should be much increased by repeating affinity capturing before and after proteolysis. In fact, glycoproteins purified on lectin columns often contain significant amounts of nonglycosylated proteins and nontarget glycoproteins, probably as a result of indirect interaction with the immobilized lectin: for example, immobilized lectin–target glycoproteins–nontarget (glyco)proteins. Inclusion of endogenous lectins is also possible, and these are likely to cross-link various glycoconjugates: for example, immobilized lectin–target glycoproteins–endogenous lectins–nontarget (glyco)proteins. If such nontarget components are analyzed, an inappropriate conclusion will be drawn. (2) Glycopeptides are much easier to separate: for example, via reversed-phase HPLC. This is an essential point in achieving comprehensive analysis of proteins. (3) It is expected that the protease treatment will reduce the undesirable action of hydrolytic enzymes such as glycosidases and proteases possibly included in the extracts.

For proteolysis, use of lysylendopeptidase[13] is highly recommended for the following reasons: (1) The enzyme is superior to any other proteases in its rigorous specificity, and thus it is highly expected that there is no cleavage other than after Lys (except that N-terminal peptides are cleaved after signal peptides). (2) The enzyme is fully active in the presence of 4 M urea or 0.1% sodium dodecyl sulfate (SDS). Therefore, almost complete hydrolysis of the bond after a Lys residue is assured. As described below, this helps greatly in gene assignment. (3) We can also expect to recover glycopeptides having average of 20 amino acids, considering the natural abundance of Lys. This size is adequate for analysis with conventional protein sequencers. With this peptide size, we can also expect approximately one glycosylation site per peptide.

In the second part of the glyco-catch method, the glycopeptides recaptured above are separated by reversed-phase chromatography and all of the relevant fractions are subjected to sequence analysis. With sequencer performance at present, at least a few picomoles of peptides are required for successful determination of certain amino acids. Therefore, drastic improvement is necessary in both sensitivity and throughput. In this regard, application of a concurrent liquid chromatography (LC)/MS/MS system seems to be most fruitful for future glycoproteomics. At the moment, however, there remain some problems for its direct application to this glyco-catch method. Therefore, in this review we describe the present procedure, which uses a protein sequencer to obtain structural information.

Notably, sequencer analysis in glycoproteomics is largely different from conventional analysis in that sequence information may be only partial and not necessarily from the N-terminus. In fact, even if complex organisms such as the mouse are analyzed, determination of six to eight amino acid residues is usually adequate for unambiguous assignment of the corresponding genes. That is, structural complexities made up of these amino acid numbers (6.4×10^7 to 2.6×10^9) are similar in number to the genome sizes of *Caenorhabditis elegans* (1.0×10^8 bp) and humans (3×10^9 bp). If reliable complementary DNA (cDNA) sequences are available, searching for unnecessary intron and junk regions can be omitted. Another distinct benefit is that one needs not purify every glycopeptide, and thus mixed sequencing is possible. This is particularly true when the SQMATCH program (described below) is used for the database search. On the other hand, identification of relatively minor components becomes more difficult in the presence of major components.

For gene assignment, the SQMATCH program, which was recently developed at the National Institute of Genetics of Japan (http://www.genome.ad.jp/SOSui/index.htm), is used as well as the more common T-BLASTN.[14] Each of them has both merits and demerits, but they are complementary. Because T-BLASTN was originally developed for homology search of databases, it is not appropriate for the purpose of search by using query sequences composed of a relatively small number of residues (e.g., <10 amino acids), such as those obtained by the glyco-catch procedure outlined above. In addition, only one candidate amino acid residue is accepted at each position of query sequences, even if the experimental sequence data clearly show the presence of more than one peptide. However, T-BLASTN tolerates some discrepancies between query and hit sequences, even though the scores become low. On the other hand, the SQMATCH program was developed with the aim of searching for genes corresponding to relatively short query sequences. Because no mismatch is permitted, there is no score-like notion as there is in T-BLASTN. Another favorable feature of SQMATCH is that one may input

multiple amino acids at each position, thus making it possible to assign multiple genes at once. In this regard, SQMATCH is preferable for the purpose of gene assignment by the glyco-catch method, although the present version of the program is not yet convenient because even only one mismatch results in rejection of a hit.

To make gene assignment more successful, skilled technique is required upon inquiry. For example, (1) because Lys at the −1 position is assured by the rigorous specificity of lysylendopeptidase (except when captured glycopeptides are derived from the N-terminus of mature proteins), the presumption of this Lys position significantly contributes to gene assignment. (2) Peptides captured by the glyco-catch method must contain at least one glycosylation site (e.g., Asn-X-Ser/Thr for *N*-glycosylation sites, and Ser/Thr clusters or so-called "mucin box" for *O*-glycosylation sites). At cycles corresponding to such glycosylation sites (i.e., Asn, Ser, or Thr), no phenylthiohydantoin (PTH) amino acids should be detected. Such presumptive input is advisable for increasing efficiency. However, to qualify the hits, thorough consideration is necessary, as listed below.

Qualification of "Hit" Genes

Only genes that satisfy all of the following criteria should be qualified:

1. Corresponding peptides are preceded by Lys (otherwise, they should follow a signal sequence).
2. Corresponding peptides contain at least one glycosylation site.
3. Corresponding peptides are eluted in reversed-phase chromatography with reasonable retention time based on their chemical nature (i.e., length and hydrophobicity).
4. All amino acids identifiable by sequence analysis should be detected as PTH derivatives with reasonable yields.
5. Assigned genes encode either a hydrophobic signal sequence for secretion (i.e., in the case of secreted glycoproteins) or at least one hydrophobic segment for membrane anchoring (i.e., membrane-anchored glycoproteins).
6. In the case that assigned genes encode membrane-anchored proteins, assumed glycosylation sites should be located in extracellular regions. For prediction in membranous regions, the SOSUI program was used in this study (http://www.genome.ad.jp/SOSui/index.html).

APPLICATION TO MOUSE COMPLEX TYPE *N*-GLYCOPROTEINS

The proposed glyco-catch method is widely applicable to various types of glycoproteins produced by organisms whose genome sequences are known. In any case, however, selection of appropriate lectin columns is most important. In addition, sufficient knowledge about lectin properties (e.g., requirement for metal ions and thiol reagents, physicochemical stability, oligomer structures, sugar-binding specificity and affinity, etc.) is required from a practical viewpoint. In this section, application of the glyco-catch method to mouse glycoproteins by using galectin-1 as a probe is described. Galectin-1 represents a major galectin in mammals, showing wide tissue distribution (liver, lung, spleen, thymus, heart, muscle, brain, etc.) and having diverse functions in a variety of biological phenomena such as development and immunity.[15] It is a prototype galectin that usually forms a noncovalent dimer consisting of two identical 14.5-kDa subunits. Most critically, the lectin has high selectivity for complex type *N*-glycans (for detailed sugar-binding specificity of galectins, see our recent review[7]). As a result of glyco-catch experiments targeting both soluble (secreted) and detergent-soluble (membrane-anchored) glycoproteins, 19 genes possibly representing major liver complex type *N*-glycoproteins were identified together with 34 *N*-glycosylation sites. All of the thus-identified genes were found to encode either secreted or membrane-anchored proteins having a hydrophobic N-terminal or internal signal sequence(s).

Preparation of Galectin-1 Column

A previously described oxidation-resistant mutant of galectin-1 designated C2S (active cysteine at position 2 substituted with serine)[16] was used. The coding region of C2S was recloned into a pET21a (Novagen, Madison, WI) vector via polymerase chain reaction (PCR) cloning using a previous cDNA cloned into pUC540(Kan^R) as a template[17] and the pCRII vector (Invitrogen, Carlsbad, CA). The galectin protein was expressed in BL21(DE3) *Escherichia coli* as a host cell and purified by affinity chromatography on an asialofetuin-agarose column (bed volume 10 mL) as described previously.[11,16,17] The purified galectin-1 (C2S) was immobilized on an *N*-hydroxysuccinimide (NHS)-activated Sepharose 4 FastFlow column (Pharmacia, Uppsala, Sweden) at a concentration of 5 mg/mL gel. Available ligand content (moles) was assessed by a reinforced FAC system using pyridylaminated lacto-*N*-fucopentaose I, Fucα1-2Galβ1-3GlcNAcβ1-3Galβ1-4Glc.[6]

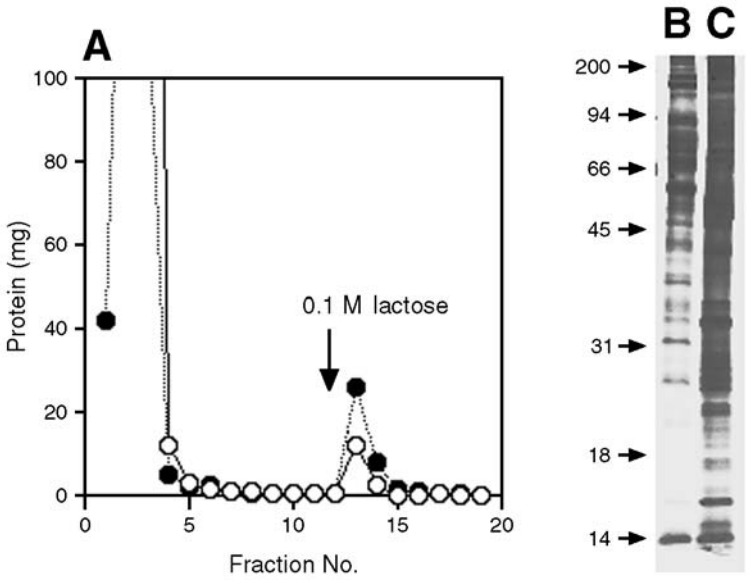

FIGURE 3

Purification of mouse liver glycoproteins by affinity chromatography on a galectin-1 (C2S) column. **A:** Chromatograms obtained for soluble (solid line, open circles) and detergent-soluble (broken line, filled circles) fractions. An arrow indicates the starting position of elution with 0.1 M lactose. **B, C:** Results of analysis of purified glycoproteins by SDS-PAGE (B, soluble fraction; C, detergent-soluble fraction). Glycoproteins eluted from the galectin-1 (C2S) column in panel A were separated by SDS-PAGE, and protein was stained with silver. Positions of marker proteins of various sizes (kDa) are indicated by arrowheads.

Preparation of Soluble and Detergent-Soluble Glycoproteins

Mouse liver (10 g, wet weight) was homogenized in 50 mL of ice-cold extraction buffer (MEPBS; 4 mM β-mercaptoethanol, 2 mM EDTA, 20 mM sodium phosphate [pH 7.2], 150 mM NaCl) by using a Polytron homogenizer (Kinematica, Lucerne, Switzerland). The homogenate was centrifuged (15,000 rpm, 4°C, 25 min), and the derived supernatant (soluble fraction) was applied to the galectin-1 (C2S) column described above (bed volume 10 mL), which had previously been equilibrated with MEPBS at 4°C. After extensive washing of the column with the same buffer, the adsorbed glycoproteins were eluted with MEPBS containing 20 mM lactose. The fraction volume was 10 mL throughout chromatography.

Insoluble glycoproteins included in the above precipitate were solubilized by gentle shaking (at 4°C for 20 min) with 25 mL of MEPBS containing 1% (w/v) Triton X-100. Solubilized protein was obtained as a supernatant (detergent-soluble fraction) by centrifugation under the same conditions as above, and was applied to another galectin-1 (C2S) column (bed volume 10 mL), which had previously been equilibrated with MEPBS containing 0.1% Triton X-100 at 4°C. Other purification procedures were the same as described for the soluble fraction except that 0.1% Triton X-100 was included throughout chromatography. For both soluble and detergent-soluble fractions, the protein concentration was determined by using a Bio-Rad Protein Assay kit, whereas protein purity was analyzed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) using 14% gels followed by silver staining with a Wako Silver Stain kit (Tokyo, Japan).

Yields of the thus-purified glycoproteins after galectin-1 (C2S) affinity chromatography for the soluble and detergent-soluble fractions were 23 mg and 15 mg, respectively. SDS-PAGE analysis showed that both fractions contained a wide range of proteins but that the overall compositions were apparently distinct between them (Fig. 3B). The latter observation confirms that extraction was performed properly.

Generation and Recapture of Glycopeptides

Glycoproteins obtained as described above were precipitated with two volumes of ethanol (−20°C) and centrifuged (9000 × *g,* 4°C, 20 min). The precipitates were 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 denaturing solution with the same volume of 0.1 M $NaHCO_3$ (pH 8.5), lysylendopeptidase (*Achromobacter* protease I;[13] Wako) at 1/100 (w/w) relative to glycoproteins was added; then digestion was conducted for 16 h at 37°C. To inactivate the protease, *p*-aminoethylbenzenesulfonyl fluoride (Wako) was added to give a final concentration of 0.1 mM, then the incubation was continued for another 1 h at 37°C. After dilution of the digest with the same volume of equilibration buffer (MEPBS), the digest was applied to the same galectin-1 (C2S) column used above, with a reduced bed volume (5 mL). After extensive washing of the column, adsorbed glycopeptides were eluted with MEPBS containing 20 mM lactose. The fraction volume was 5 mL throughout chromatography.

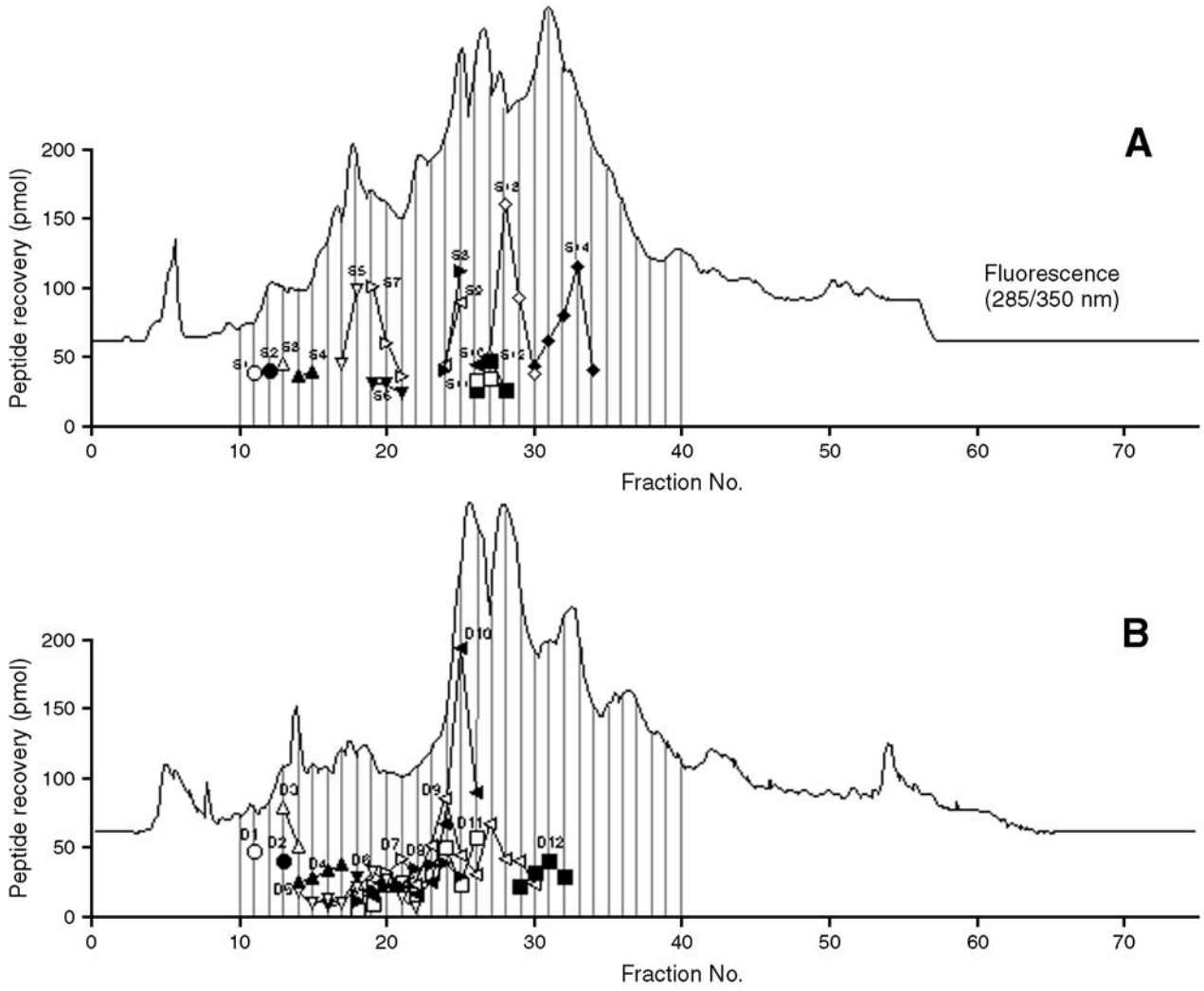

FIGURE 4

Separation of the glycopeptides derived from the soluble fraction **(A)** and the detergent-soluble fraction **(B)** by reversed-phase chromatography. Glycopeptides recaptured by the glyco-catch method were separated by HPLC on a trimethylsilyl Develosil UG-5 column (Nomura Chemicals; 4.6 × 150 mm) by a conventional linear gradient elution system using 0.1% trifluoroacetic acid/acetonitrile (5% to 45%). Elution of peptides was monitored by ultraviolet absorbance (210 nm) and by fluorescence based on tryptophan (excitation and emission wavelengths 285 and 350 nm, respectively), but only the results of the latter are shown in the figure. All consecutive fractions 11–40 (vertical lines) were subjected to structural analysis with a Shimadzu PPSQ-21 protein sequencer. Recoveries of assigned glycopeptides (designated S1–S14 and D1–D12; for details see Tables 1 and 2, respectively) are also included in the figure.

Separation of Glycopeptides and Gene Assignment

The recaptured glycopeptide fractions above (20 mL each) were concentrated by ultrafiltration using a Centriplus YM-3 (Millipore, Bedford, MA; exclusion size 3000 Da) and fractionated by reversed-phase chromatography on a trimethylsilyl (TMS) Develosil UG-5 column (Nomura Chemicals, Tokyo, Japan; 4.6 × 150 mm). Glycopeptides were eluted with a linear gradient of increasing concentrations of acetonitrile in 0.1% (v/v) trifluoroacetic acid—5% to 15% (0–5 min) then 15% to 45% (5–55 min)—at a flow rate of 1.0 mL/min. Elution of glycopeptides was monitored by both ultraviolet absorbance at 210 nm (A_{210}) and fluorescence based on tryptophan (excitation and emission wavelengths 285 and 350 nm, respectively) (only the results of fluorescence monitoring are shown in Fig. 4). Frac-

TABLE 1

Assigned Peptides Derived from the Soluble Fraction

ID	Analyzed sequence[a]	Fraction(s)	Protein name or designation[b]	Accession no.
S1	K. YL**N**ETQQLTQK.	11	Mug	M65736-1
S2	K. NLFL**N**HSETASAK.	12	Haptoglobin	M96827-1
S3	K. HSIEHFN**N**NT DHSHL.....K.	13	HMW prekininogen	D84435-1
S4	K. **N**ATSYPPMCS QDAG.....K.	14, 15	Carboxylerastase	M57960
S5	K. V**N**LSFPSAQ SLPASDTHLK.	17, 18	α2M	M93264-1
S6	K. **N**ISFACNPGF FL**N**GTSSSK.	19–21	Apolipoprotein H	Y11356-1
S7	Q. NLINDYVS**N**Q TQGMIK. ← signal sequence	19–21	Contraspin or Spi-2/ACT	X56786-1 X69832-1
S8	K. NPEHA**N**FTIG EPIT**N**ETLS.....K.	24, 25	Agp	M27008-1
S9	K. SLGEVNFTRT AEALESQEL.....K.	24, 25	α2M	M93264-1
S10	K. ATLSVLV**N**AS TGHLLLPIEN.....K.	26	VE cadherin-2	Y08715-1
S11	K. VPFIFNINPA TT**N**FTGSCQP.....K.	26, 27	Lamp-2	J05287-1
S12	K. YTG**N**ASALLI LPDQGRMQQV.....LRK.	26–28	Contraspin	X56786-1
S13	K. SPLPTAHGRV AEVE**N**GTK. ← signal sequence	27–30	Hemopexin	BC0111246
S14	K. AFE**N**VTDLQW LILDHNLLEN SK.	30–341	Lum	AF013262-1

[a]Confirmed *N*-glycosylation sites are denoted as **N**. Residues preceding the assigned peptides were lysine except for S7, which is followed by a presumed signal sequence for secretion ending with glutamine.

[b]Designations of the listed proteins are as follows: Mug, murinoglobulin; HMW prekininogen, high-molecular-weight prekininogen; α2M, α2-macroglobulin; Spi-2, serine protease inhibitor 2; ACT, α1-antichymotrypsin; Agp, α1-acid glycoprotein; VE cadherin-2, vascular endothelial cadherin-2; Lamp-2, lysosomal-associated membrane glycoprotein-2; Lum, lumican.

tions were collected every minute, and one-third of each was subjected to automated analysis with a Shimadzu (Kyoto, Japan) PPSQ-21 protein sequencer. In this study, all fractions between 10 and 40 were analyzed. As a result, 14 glycopeptides from the soluble fraction (designated S1–S14) and 12 glycopeptides from the detergent-soluble fraction (designated D1–D12) were successfully analyzed and unambiguously assigned to their corresponding cDNAs in the mouse cDNA database[18] (Tables 1, 2). However, it was not possible to assign one peptide, designated S7, because there were two candidates, X56786-1 (contraspin) and X69832-1 (serine protease inhibitor, Spi-2). They are highly homologous to each other and thus encode the same peptide sequence, which corresponds to S7.

Features of the thus-assigned 14 soluble and 12 detergent-soluble peptides are summarized as follows: (1) All of them were preceded by Lys except for S7 and D2, both of which were derived from the N-terminus of mature proteins, contraspin (X-56786-1) and lamp-2 (J05287-1), respectively. (2) All of the assigned peptides contained either one or two possible *N*-glycosylation sites, "Asn-X-Ser/Thr." The former include S1–S5, S7, S9–S12, S13, and S14 from the soluble fraction and D1–D4, D6, D10, and D12 from the detergent-soluble fraction, whereas the latter include S6 and S8 from the soluble fraction and D5, D7–D9, and D11 from the detergent-soluble fraction. At all of these potential glycosylation sites except for the first Asn of peptide D7, PTH-Asn was not significantly detected. Therefore, actual glycosylation was confirmed for 16 sites from S1–S14 and 16 sites from D1–D12. On the other hand, PTH derivatives of the first Asn (i.e., the seventh position) of D7 were unambiguously detected; therefore, it was not glycosylated despite its potential. (3) Thirteen cDNAs (i.e., 13 genes) have been identified in the database as those encoding glycopeptides S1–S14, whereas seven cDNAs (six genes) have been found to encode glycopeptides D1–D12 (note that organic anion transporting peptides [Oatp] 1 and 2 are the products of alternative splicing of the same gene[19]).

TABLE 2

Assigned Peptides Derived from the Detergent-Soluble Fraction

ID	Analyzed sequence[a]	Fraction(s)	Protein designation[b]	Accession no.
D1	K. VQPF**N**VTK.	11	Lamp-2/Lgp110	J05287-1
D2	A. LIV**N**LTDSK. ← signal sequence	13	Lamp-2/Lgp110	J05287-1
D3	K. ANIQFGE**N**GT TISAVTNK.	13, 14	Lgp85/Limp-II	AB008553-1
D4	K. SVGTGTNMVF Q**N**CSCIGSSG.....K.	14–17	Oatp 1 Oatp 2	AB031813-1 AB031814-1
D5	K. NITVLEPVTQ PFLQVT**N**TTV K.	14–22	MHVR-1/Bgp/CEA	M77196-1
D6	K. IP**N**NTQWITW SPEGHK.	16–18	Dpp IV/CD26	U12629-1
D7	K. TVTRAFNISP NDTSSGSCGI.....K.	18–22	Lamp-1/Lgp-A	M32015-1
D8	K. NTTRYSVQHM YFTY**N**LSDTE.....K.	18–25	Lamp-1/Lgp-A	M32015-1
D9	K. NVTVVLRDAT IQAYLSSG**N**F SK.	18–30	Lamp-1/Lgp-A	M32015-1
D10	K. VPFIFNINPA TT**N**FTGSCQP.....K.	18–26	Lamp-2/Lgp-B	J05287-1
D11	K. EVNVYMYLAN GSAF**N**ISNK.	18–26	Lamp-2/Lgp-B	J05287-1
D12	K. LSEGNRTLTL LNVTR**N**DTGP.....K.	29–32	MHVR-1/Bgp/CEA	M77196-1

[a]**N** represents a confirmed *N*-glycosylation site, whereas N represents a confirmed nonglycosylation site.

[b]Designations of the listed proteins are as follows: Lamp-1/2, lysosomal-associated membrane glycoprotein-1/2; Lgp110/85, lysosomal glycoprotein 110/85; Limp-II, lysosomal integral membrane protein II; Oatp 1/2, organic-anion-transporting polypeptide 1/2; MHVR-1, mouse hepatitis virus receptor 1; Bgp, biliary glycoprotein; CEA, carcinoembryonic antigen; Dpp IV, dipeptidyl peptidase IV; CD26, cluster of differentiation 26.

For all of the assigned glycopeptides, full-length cDNAs have already been isolated and thus registered in the mouse cDNA database except for Spi-2 (X69832-1), for which only a partial sequence has been determined.[20] (4) The identified genes for the soluble-fraction proteins encoded 11 secreted and two membrane-anchored proteins (both type I membrane proteins), whereas those for the detergent-soluble proteins encoded seven membrane-anchored proteins (four type I, one type II, and two multiple membrane-anchored proteins). The former group consisted of typical secreted proteins: hydrolases (carboxylesterase, high-molecular-weight prekininogen), protease inhibitors (α2M, contraspin, Spi-2, murinoglobulin), extracellular matrix proteins (lumican), and others (α-1 acid glycoprotein, hemopexin, haptoglobin, apolipoprotein H). Inclusion of the two membrane proteins in the soluble fraction (i.e., lamp-2 and VE-cadherin) may be attributed to either partial proteolysis during the extraction procedure or alternative splicing, which has not yet been characterized.

In total, we identified 26 *N*-glycosylation sites of 26 glycopeptides encoded by 20 cDNAs derived from 19 genes. The features of these assigned glycoproteins are summarized in Table 3, and overall structures depicting both potential and confirmed *N*-glycosylation sites are shown schematically in Figure 5.

PERSPECTIVE

Undoubtedly, glycomics represents a most important field of post-genome sciences to be explored in the 21st century. The glyco-catch procedure described in this review is expected to constitute a core strategy for glycomics targeting glycoproteins. The principal aim of glycoproteomics is to understand various complex life systems by applying modern biotechnologies to the field of glycobiology under the concept of "genome-proteome-glycome."[3–5] To achieve this goal, however, much remains to be improved in the present technology. In particular, achievement of higher throughput and performance with higher sensitivity is essential. Though still preliminary, such an attempt utilizing a glyco-catch/LC/MS/MS strategy is already in progress (Kaji and Isobe, personal communication). In addition, application of the glyco-catch method to the analysis of *O*-glycoproteins remains to be improved,

TABLE 3

cDNAs Identified by the Present Glyco-Catch Analysis

Accession no.	Protein name (designation)	Protein type[a]	No. AA residues	Assigned region (ID)	*N*-glycos. site	Ref(s).
SOLUBLE FRACTION						
AF013262-1	Lumican (Lum)	Secreted	338	85–106 (S14)	88	23
M57960-1	Carboxylesterase	Secreted	554	79–95 (S4)	79	24
D84435-1	HMW prekininogen	Secreted	661	161–175 (S3)	168	25
M27008-1	α-1 acid glycoprotein (Agp)	Secreted	707	20–38 (S8)	25 34	26
M65736-1	Murinoglobulin (Mug)	Secreted	1476	991–1001 (S1)	993	27, 28
M93264-1	α-2 macroglobulin (α2M)	Secreted	1496	567–585 (S5) 876–894 (S9)	568 881	29
M96827-1	Haptoglobin	Secreted	347	144–156 (S2)	148	29
BC011246	Hemopexin	Secreted	459	24–41 (S13)	38	30
X56786-1	Contraspin	Secreted	418	177–192 (S7) 267–297 (S12)	185 270	31
X69832-1	Serine proteinase inhibitor (Spi-2, ACT)	Type I	275	32–47 (S7)	40	19
Y08715-1	Vascular endotherial cadherin-2	Secreted	1180	575–594 (S10)	582 594	32
Y11356-1	Apolipoprotein H	Secreted	345	105–123 (S6)	105 117	33
J05287-1	Lamp-2	Type I	415	253–272 (S11)	265	34, 35
DETERGENT-SOLUBLE FRACTION						
AB031813-1	Organic anion transporter 1 (Oatp 1)	Multi	670	472–501 (D4)	483	36
AB031814-1	Organic anion transporter 2 (Oatp 2)	Multi	670	472–502 (D4)	483	18
M32015-1	Lamp-1	Type I	406	101–120 (D8)	101 115	37
				159–178 (D9)	159 177	
				242–261 (D7)	252	
J05287-1	Lamp-2	Type I	415	26–34 (D2)	29	34,35
				253–272 (D10)	265	
				303–321 (D11)	312 317	
				357–364 (D1)	362 367	
M77196-1	Hepatitis virus receptor (MHVR 1)	Type I	458	195-214 (D12)	199	38
	Biliary glycoprotein (Ggp)				206	39
	Carcinoenbryonic antigen (CAE)				210	40
				317–336 (D5)	317 333	
U12620-1	Dipeptidyl peptidase IV (Dpp IV/CD26)	Type II	760	142–157 (D6)	144	41
AB008553-1	Lgp85/Limp-II	Type I	478	98–115 (D3)	105	42

[a]Protein types are based on *in silico* prediction by SOSUI (http://www.genome.ad.jp/SOSui/index.html) for membrane-integral hydrophobic regions, i.e., secreted proteins (Secreted), type I (Type I) and type II (Type II) membrane proteins, multiple-membrane anchored proteins (Multi), and cytoplasmic proteins (no case in this study).

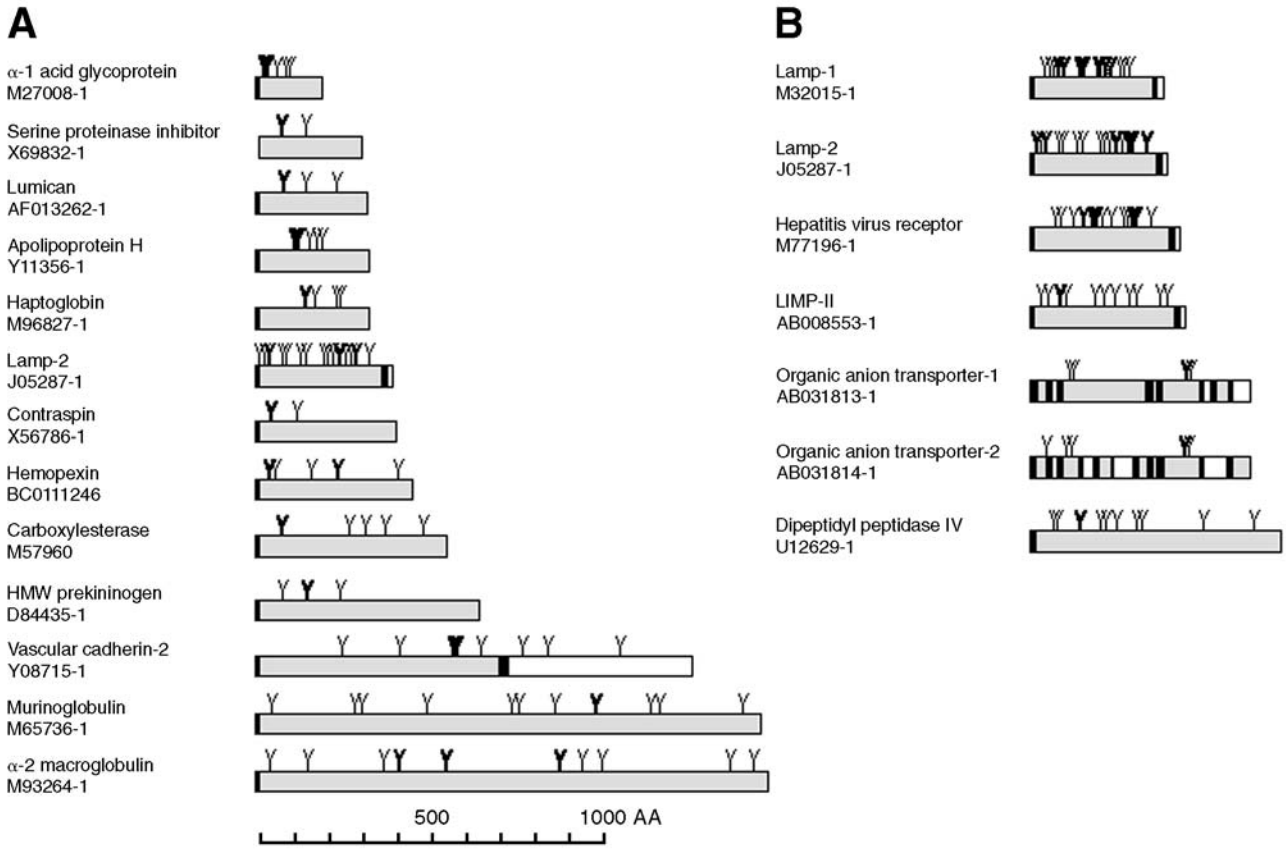

FIGURE 5

Schematic representation of identified glycoproteins derived from the soluble **(A)** and detergent-soluble **(B)** fractions. Potential *N*-glycosylation sites are shown by thin "Y" characters, and the experimentally confirmed *N*-glycosylation sites are shown by bold "**Y**" characters. Descriptive names for individual gene products are shown with their accession numbers, which have been registered in the GenBank/EMBL/DDBJ databases.

as discussed previously.[10] If such improvements are attained, extensive glycomics investigation targeting various types of cells and individuals can be conducted more rapidly and efficiently, through various approaches. For example, glycomics can be applied to medical sciences in a manner similar to single nucleotide polymorphism (SNP) analysis. Almost all extracellular proteins, covering every kind of cell, are modified by glycans[21] and, hence, many genetic diseases may be related to a lack of or abnormality of genes responsible for glycan synthesis and metabolism.[22,23] Therefore, glycomic projects at various levels should contribute not only to pure science (i.e., for the elucidation of biological functions of glycans and ultimately for deciphering the "glycocode") but also to applied science (i.e., medical science, "glycoindustries", etc.). In this context, construction of a useful, versatile glycoproteome database should become more and more a key issue for glycomics in the future.

ACKNOWLEDGMENTS

This work was supported in part by grants from Grants-in-Aid for Scientific Research on Priority Area "Genome Science" (no. 13202058 to J.H.) and Grants-in-Aid for Scientific Research (nos. 12680617 to J.H. and 11771453 to K.K.) of the Ministry of Education, Science, Sports, and Culture of Japan and by the Mizutani Foundation for Glycoscience.

REFERENCES

1. Varki A, Cummings R, Freeze H, Hart G, Marth J (eds). *Essentials of Glycobiology*. New York: Cold Spring Harbor Laboratory Press, 1999.
2. Laine RA. A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05×10^{12} structures for a reducing hexasaccharide: the isomer barrier to development of single-method saccharide sequencing or synthesis system. *Glycobiology* 1994;4: 759–767.
3. Hirabayashi J, Kasai K. Invitation to the glycopeptide glycome project of *C. elegans*. *Glycoconjugate J* 1999; 16:S33.

4. Hirabayashi J, Kasai K. Glycomics, coming of age! *Trends Glycosci Glycotechnol* 2000;12:1–5.
5. Hirabayashi J, Arata Y, Kasai K. Glycome project: concept, strategy and preliminary application to *C. elegans*. *Proteomics* 2001;1:295–303.
6. Hirabayashi J, Hashidate T, Arata Y, et al. Oligosaccharide specificity of galectins: a search by frontal affinity chromatography. *Biochim Biophys Acta* 2002;1572:232–254.
7. Hirabayashi J, Kaji H, Isobe T, Kasai K. Affinity capturing and gene assignment of soluble glycoproteins produced by the nematode *Caenorhabditis elegans*. *J Biochem (Tokyo)* 2002;132:103–114.
8. Takahashi N, Wada Y, Awaya J, Kurono M, Tomiya N. Two-dimensional elution map of GalNAc-containing *N*-linked oligosaccharides. *Anal Biochem* 1993;208:96–109.
9. Natsuka S, Hase S. Analysis of *N*- and *O*-glycans by pyridylamination. In Hounsell EF (ed): *Methods in Molecular Biology,* vol 76. Totowa, NJ: Humana Press, 1998: 101–113.
10. Hirabayashi J, Kasai K. Separation technologies for glycomics. *J Chromatogr B* 2002;771:67–87.
11. Hirabayashi J, Arata Y, Kasai Y. Reinforcement of frontal affinity chromatography for effective analysis of lectin–oligosaccharide interactions. *J Chromatogr A* 2000;890: 261–271.
12. Arata Y, Hirabayashi J, Kasai K. Sugar binding properties of the two lectin domains of the tandem repeat-type galectin LEC-1 (N32) of *Caenorhabditis elegans*: detailed analysis by an improved frontal affinity chromatography method. *J Biol Chem* 2001;276:3068–3077.
13. Sakiyama F, Masaki T. Lysyl endopeptidase of *Achromobacter lyticus*. *Methods Enzymol* 1994;244:126–137.
14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215: 403–410.
15. Hirabayashi J (ed). Recent topics on galectins. *Trends Glycosci Glycotechnol* 1997;9:1–180.
16. Hirabayashi J, Kasai K. Effect of amino acid substitution by site-directed mutagenesis on the carbohydrate recognition and stability of human 14-kDa β-galactoside binding lectin. *J Biol Chem* 1990;266:23648–23653.
17. Hirabayashi J, Ayaki H, Soma G, Kasai K. Production and purification of a recombinant human 14 kDa β-galactoside-binding lectin. *FEBS Lett* 1989;250:161–165.
18. Kawai J, Shinagawa A, Shibata V, et al. Functional annotation of a full-length mouse cDNA collection. *Nature* 2001;409:685–690.
19. Ogura K, Choudhuri S, Klaassen CD. Genomic organization and tissue-specific expression of splice variants of mouse organic anion transporting polypeptide 2. *Biochem Biophys Res Commun* 2001;281:431–439.
20. Inglis JD, Lee M, Davidson DR, Hill RE. Isolation of two cDNAs encoding novel α 1-antichymotrypsin-like proteins in a murine chondrocytic cell line. *Gene* 1991; 106:213–220.
21. Apweiler R, Hermjankob H, Sharon N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta* 1473;1999:4–8.
22. Stanley P, Ioffe E. Glycosyltransferase mutants: key to new insights in glycobiology. *FASEB J* 1995;9:1436–1444.
23. Esko JD. Animal cell mutants defective in heparan sulfate polymerization. *Adv Exp Med Biol* 1992;313:97–106.
24. Ying S, Shiraishi A, Kao CW, et al. Characterization and expression of the mouse lumican gene. *J Biol Chem* 1997;272:30306–30313.
25. Ovnic M, Tepperman K, Medda S, et al. Characterization of a murine cDNA encoding a member of the carboxylesterase multigene family. *Genomics* 1991;9:344–354.
26. Takano M, Kondo J, Yayama K, Otani M, Sano K, Okamoto H. Molecular cloning of cDNAs for mouse low-molecular-weight and high-molecular-weight prekininogens. *Biochim Biophys Acta* 1997;1352:222–230.
27. Lee SC, Chang CJ, Lee YM, Lei HY, Lai MY, Chen DS. Molecular cloning of cDNAs corresponding to two genes of α-1 acid glycoprotein and characterization of two alleles of AGP-1 in the mouse. *DNA* 1989;8:245–251.
28. Overbergh L, Torrekens S, Van Leuven F, Van den Berghe H. Molecular characterization of the murinoglobulins. *J Biol Chem* 1991;266:16903–16910.
29. van Leuven F, Torrekens S, Overbergh L, Lorent K, de Strooper B, van den Berghe H. The primary sequence and the subunit structure of mouse alpha-2-macroglobulin, deduced from protein sequencing of the isolated subunits and from molecular cloning of the cDNA. *Eur J Biochem* 1992;210:319–327.
30. Yang F, Linehan LA, Friedrichs WE, Lalley PA, Sakaguchi AY, Bowman BH. Characterization of the mouse haptoglobin gene. *Genomics* 1993;18:374–380.
31. Nikkila H, Gitlin JD, Muller-Eberhard U. Rat hemopexin. Molecular cloning, primary structural characterization, and analysis of gene expression. *Biochemistry* 1991;30:823–829.
32. Suzuki Y, Yamamoto K, Shinohara H. Molecular cloning and sequence analysis of full-length cDNA coding for mouse contraspin. *J Biochem (Tokyo)* 1990;108:344–346.
33. Telo' P, Breviario F, Huber P, Panzeri C, Dejana E. Identification of a novel cadherin (vascular endothelial cadherin-2) located at intercellular junctions in endothelial cells. *J Biol Chem* 1998;273:17565–17572.
34. Nonaka M, Matsuda Y, Shiroishi T, Moriwaki K, Nonaka M, Natsuume-Sakai S. Molecular cloning of mouse β2-glycoprotein I and mapping of the gene to chromosome. *Genomics* 1992;13:1082–1087.
35. Cha Y, Holland SM, August JT. The cDNA sequence of mouse LAMP-2. Evidence for two classes of lysosomal membrane glycoproteins. *J Biol Chem* 1990;265:5008–5013.
36. Granger BL, Green SA, Gabel CA, Howe CL, Mellman I, Helenius A. Characterization and cloning of lgp110, a lysosomal membrane glycoprotein from mouse and rat cells. *J Biol Chem* 1990;265:12036–12043.
37. Hagenbuch B, Adler ID, Schmid TE. Molecular cloning and functional characterization of the mouse organic-anion-transporting polypeptide 1 (Oatp1) and mapping of the gene to chromosome X. *Biochem J* 2000; 345:115–120.
38. Chen JW, Cha Y, Yuksel KU, Gracy RW, August JT. Isolation and sequencing of a cDNA clone encoding lysosomal membrane glycoprotein mouse LAMP-1. Sequence similarity to proteins bearing onco-differentiation antigens. *J Biol Chem* 1988;263:8754–8758.
39. Dveksler GS, Pensiero MN, Cardellichio CB, et al. Cloning of the mouse hepatitis virus (MHV) receptor: expression in human and hamster cell lines confers susceptibility to MHV. *J Virol* 1991;65:6881–6891.
40. McCuaig K, Rosenberg M, Nedellec P, Turbide C, Beauchemin N. Expression of the Bgp gene and characterization of mouse colon biliary glycoprotein isoforms. *Gene* 1993;127:173–183.

41. Beauchemin N, Turbide C, Afar D, et al. A mouse analogue of the human carcinoembryonic antigen. *Cancer Res* 1989;49:2017–2021.
42. Marguet D, Bernard AM, Vivier I, Darmoul D, Naquet P, Pierres M. cDNA cloning for mouse thymocyte-activating molecule. A multifunctional ecto-dipeptidyl peptidase IV (CD26) included in a subgroup of serine proteases. *J Biol Chem* 1992;267:2200–2208.
43. Tabuchi N, Akasaki K, Sasaki T, Kanda N, Tsuji H. Identification and characterization of a major lysosomal membrane glycoprotein, LGP85/LIMP II in mouse liver. *J Biochem (Tokyo)* 1997;122:756–763.