ELSEVIER

Journal of Chromatography B, 771 (2002) 67–87

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Review

Separation technologies for glycomics

Jun Hirabayashi*, Ken-ichi Kasai

Department of Biological Chemistry, Faculty of Pharmaceutical Sciences, Teikyo University, Sagamiko, Kanagawa 199-0195, Japan

Abstract

Progress in genome projects has provided us with fundamentals on genetic information; however, the functions of a large number of genes remain to be elucidated. To understand the in vivo functions of eukaryotic genes, it is essential to grasp the features of their post-translational modifications. Among them, protein glycosylation is a central issue to be discussed, considering the predominant roles of glycoproteins in cell–cell and cell–substratum recognition events in multicellular organisms. In this context, it is necessary to establish a core strategy for analyzing glycosylated proteins under the concept of the ''glycome'' [Trends Glycosci. Glycotechnol. 12 (2000) 1]. Though the term glycome should be defined, in analogy to the genome and proteome, as ''a whole set of glycans produced in a single organism'', here we propose a glycome project specifically focusing on glycoproteins. Principal objectives in the project are to identify: (1) which genes encode glycoproteins (i.e. genome information); (2) which sites among potential glycosylation sites are actually glycosylated (i.e. glycosylation site information); (3) what are the structures of glycans (i.e. structural information); and (4) what are the effects (functions) of glycosylation (functional information). For these purposes, two affinity technologies have been introduced. One is named the ''glyco-catch method'' to identify genes encoding glycoproteins [Proteomics 1 (2001) 295], and the other is the recently reinforced ''frontal affinity chromatography'' [J. Chromatogr. A 890 (2000) 261]. By the former method, genes that encode glycoproteins as well as glycosylation sites are systematically identified by the efficient combination of conventional lectin-affinity chromatography and contemporary in silico database searching. The following three actions have been devised for rapid and systematic characterization of glycans: (1) mass spectrometry to acquire exact mass information; (2) 2-D/3-D mapping to obtain refined chemical information; and (3) reinforced frontal affinity chromatography to determine affinity constants (K_a-values) for a set of lectins. Pyridylaminated glycans are used throughout the characterization processes. In this review, the concept and strategy of glycomic approaches are described referring to the on-going glycome project focused on the nematode *Caenorhabditis elegans*.

Keywords: Reviews; Glycomics

Contents

*Corresponding author. Tel.: +81-426-85-3741; fax: +81-426-85-3742.
E-mail address: j-hira@pharm.teikyo-u.ac.jp (J. Hirabayashi).

1570-0232/02/$ – see front matter
PII: S1570-0232(02)00057-0

1. Definition of "glycome/glycomics" and their purpose

"Glycome" is a conceptual word proposed at the end of 20th century to mean the whole set of glycans produced in a single organism [1–4]. Apparently, this word is analogous to genome and proteome, meaning a whole set of genes and proteins, respectively. Similarly, "glycomics" is defined as research on the glycome in analogy to genomics and proteomics. Though in the literature these terms appeared for the first time in the late 1990s [1], the concept seems to have long been shared by several glycobiologists [5–7]. In this context, the concept glycome should have arisen, in principle, when V.C. Washinger referred to the concept of the "proteome" in 1995 [8]. Importantly, he developed a novel strategy for gene-product mapping of *Mycoplasma genitalium* by combining 2-D electrophoresis originally developed by O'Farrell [9] and the recently improved matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI–TOF MS).

We have so far asserted the importance to study glycans from a genome-wide viewpoint to understand complex life systems [2]. Reasons for this assertion are as follow: *firstly*, all cells of all organisms are covered with abundant and heterogeneous glycans, of which compositions significantly reflect differences in cell types and states, e.g. species, individuals, tissues, developmental stages, malignancy, etc. However, glycans are not merely markers to characterize each cell type, but are more aggressively involved in numerous biological phenomena, such as cell development, differentiation, implantation, morphogenesis, tumor metastasis, microbe infection, etc. [10]. Notably, cultured mammalian cells with mutations in the biosynthetic pathways of protein glycosylation showed no apparent aberrant phenotypes while mice defective in the relevant genes die as embryos [11]. This fact strengthens the importance of glycans for multicellular animals to maintain "unity".

Secondly, glycans have much higher potential to exert structural diversity than nucleic acids and proteins [12]. Having sufficient diversity is essential for biologically informative molecules. Apparently, such high complexity of glycans is exclusively attributed to variation in linkage and branching events, which cannot be said of other bio-informative macromolecules. On the other hand, the number of component saccharides is relatively small, not more than Glc, GlcNAc, Man, Gal, GalNAc, L-Fuc, Xyl, L-Ara, and NeuAc. Owing to this structural diversity, the set of glycans expressed on a certain cell surface can play the role of a "bar code".

Thirdly, glycans reflect many evolutionary aspects [13]; e.g. these include: (1) the matter of chirality, (2) formose reaction, (3) glycolysis, (4) Lobry-de-Bruyn rearrangement, (5) biosynthetic features, (6) late appearance of Gal, and finally, (7) ''the origin of ribose'' is a question that has never been answered.

As depicted by recent findings [14–16], the context of glycans is fairly distinct from that of genes and proteins (genetic code). Therefore, unless we adopt a global viewpoint of ''genome–proteome–glycome'', we will never decipher the final life code, i.e. the ''glycocode''. Glycomics is what we should study in order to understand more fully such complex life systems.

2. A practical approach to glycomics: focus on glycoproteins

To undertake glycome projects, the following three actions are needed as a core strategy: (1) the whole set of glycans produced in a single organism are analyzed, (2) glycopeptides are targeted to identify genes that encode glycoproteins, and (3) glycans attached to glycoproteins are characterized by effective combination of physicochemical and biochemical criteria.

The first action is none other than the concept of the glycome itself. However, it should be mentioned that the proposed strategy for glycomics is specifically designed for glycoproteins: and hence, different approaches are necessary for glycolipids and proteoglycans. On the other hand, the other two actions mostly feature the project described here. Adoption of glycopeptides rather than released glycans as a registered unit is a key to linkage to the established genome databases. So far, complete or almost complete genome (or cDNA) sequences have been determined for several multicellular organisms, e.g. *C. elegans* [17], *Drosophila melanogaster* [18], *Arabidopsis thaliana* [19], *Mus musculus* [20], *Homo sapiens* [21]. These are also principal targets for glycomics. In eukaryotic proteins, there are three types of glycosylation, i.e. *N*-glycosylation, *O*-glycosylation, and GPI-anchoring (Table 1). For these genome-defined organisms, relevant proteome databases are already available. At the moment, however, only limited information on protein glycosylation has been annotated to each gene product. In fact, more than 90% of glycosylation data annotated to the *C. elegans* proteome database (e.g. http://www.proteome.com/databases/) are based on just in silico prediction (Table 2). By the glyco-catch method (Fig. 1) described in Section 2.1, however, genes that encode glycoproteins (gene information) as well as glycosylation sites among potential glycosylation sites (glycosylation site information) can be systematically assigned.

The third point will be a major issue, if one undertakes a glycome project of any organism; i.e. how to characterize (or ''define'') each glycan structure in the context of high-throughput proteomics. Though it is ideal for glycomics to determine all covalent structures of glycans, it is obviously impractical with the lack of an automated glycan sequencer. Rather than following such a gene/protein-type strategy, extraction of the essence of each glycan would be more practical. In this regard, we have settled on the following three actions based on different principles to characterize glycan structures (Fig. 2): (1) MS analysis to acquire exact mass values (*physical approach*; described in Section 3.1); (2) 2-D/3-D mapping of pyridylaminated derivatives to characterize chemical properties, i.e. charge, molecular size and hydrophobicity [22,23] (*chemical approach*; described in Section 3.2), and (3) reinforced frontal affinity chromatography [3,24,25] to investigate affinity to lectins (*biochemical approach*, described in Section 3.3).

By the last procedure, affinity profiling with various lectins in terms of association constants (K_a) is attained. Adoption of a set of K_a-values to a defined set of lectins is promising, since lectins are supposed to function in vivo as ''decipherers'' of complex glycans [26]. Actually, lectins have high potential to discriminate subtle differences in glycan structures (e.g. linkages and modifications). Notably, the above three procedures can be well carried out using pyridylaminated (PA) oligosaccharides. Though the PA method was originally developed as a fluorescence labeling technique for oligosaccharides by Hase et al. [27], it was subsequently used for 2-D/3-D mapping of oligosaccharides by Takahashi et al. [22,23]. Moreover, sensitivity of detection of PA-oligosaccharides by MS analysis is greatly enhanced compared with that of intact forms. As

Table 1
Types of glycosylation widely found in animal glycoproteins

Categories	Subtypes	Representative structures
N-Glycosylation		
	High-mannose type	Manα2Manα6 Manα6 Manα2Manα3 Manβ4GlcNAcβ4GlcNAcβAsn Manα2Manα2Manα3
	Hybrid type	Manα6 Manβ4GlcNAcβ4GlcNAcβAsn Galβ4GlcNAcβ2Manα3
	Complex type	Galβ4GlcNAcβ2Manα6 Manβ4GlcNAcβ4GlcNAcβAsn Galβ4GlcNAcβ2Manα3
O-Glycosylation		
	O-GalNAc	
	Core 1	Galβ3GalNAcαSer/Thr
	Core 2	GlcNAβ6 GalNAcαSer/Thr Galβ3
	Core 3	GlcNAcβ3GalNAcαSer/Thr
	Core 4	GlcNAcβ6 GalNAcαSer/Thr GlcNAcβ3
	O-GlcNAc (Nuclear protein)	
	O-Man (α-Dystroglycan)	
	O-Fuc (Notch/Fringe)	
	O-Xyl (proteoglycan)	
GPI-anchoring	Protein-CONH-CH_2CH_2-OPO_3^--2Manα2Manα6Manα4GlcNα6PI(Acyl)	

described in Section 3.3, PA-oligosaccharides are the best form for use in reinforced frontal affinity chromatography. There are some other labeling methods giving still higher sensitivity; however, only a few of them are superior to the PA-method in resolution in reversed-phase chromatography, possibly because of the small size (i.e. hydrophobicity) of the PA group, which gives rise to rapid on/off ratio upon binding to C_{18} resin.

2.1. Glyco-catch method: a basic strategy

The basic procedure of glyco-catch method is outlined in Fig. 1. The method is enabled by a novel combination of conventional lectin affinity chromatography and in silico database searching. Glycoproteins contained in crude extracts of cells and tissues are trapped by an affinity adsorbent bearing a certain type of lectin (*lectin affinity-1*). Adsorbed glycopro-

Table 2
In silico information of protein glycosylation in *C. elegans* proteome database[a]

Types of glycosylation	Total	Prediction	Experiment	(Cosmid)
N-Glycosylation	45	43 (96%)	2 (4%)	(ZK945.9, C44B12.2)
O-Glycosylation	4	3 (75%)	1 (25%)	(C44B12.2)
Glycosylation (unknown type)	17	13 (76%)	4 (24%)	(T01C8.7, C42D8.2, C04F6.1, K07H8.8)
GPI anchoring	83	83 (100%)	0 (0%)	

[a] In the proteome database [http://www.proteome.com/], *C. elegans* genes having annotation as regards glycosylation are counted, and classified into two categories based on whether the information is derived by in silico prediction or experimental evidence.

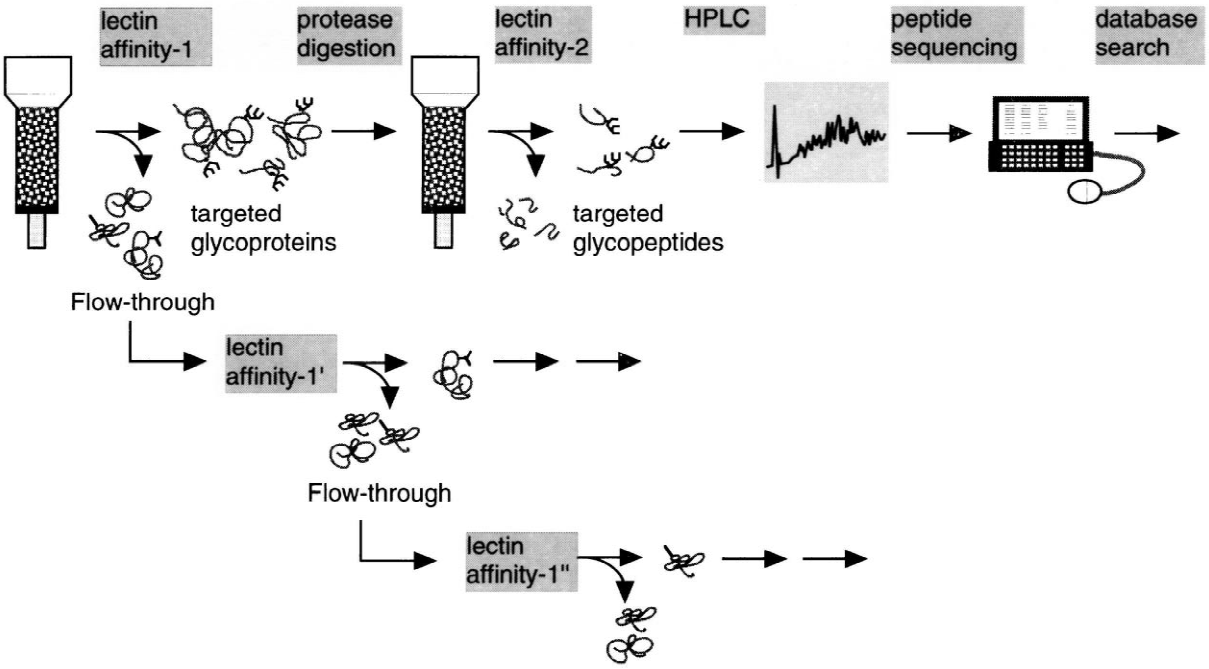


Fig. 1. Scheme for glyco-catch procedures (originally published in Ref. [3]). (1) *Lectin affinity-1*: groups of glycoproteins are captured using distinct types of lectin columns. (2) *Protease digestion*: pooled glycoproteins are extensively digested with *Achromobacter* protease I (lysylendopeptidase). (3) *Lectin affinity-2*: the resultant digest is applied to the same lectin column used in (1) to recover specifically target glycopeptides. (4) *HPLC*: recovered glycopeptides are separated by means of conventional reversed-phase chromatography. (5) *Peptide sequencing*: each fraction is subjected to analysis by a protein sequencer. (6) *Database search*: based on the derived information on partial amino acid sequences, the genome database is searched by either the BLAST-N [48] or SQMATCH (http://ftp2.ddbj.nig.ac.jp:8080/sqmatch_ja.html) programs.

teins are eluted and subjected to extensive proteolysis (*protease digestion*), and the resultant peptides are applied to the same lectin column as used in *lectin affinity-1* to select target glycopeptides (*lectin affinity-2*). Then, the obtained glycopeptides are separated by a reversed-phase HPLC (*HPLC*), and each fraction is analyzed by sequencer (*peptide sequencing*). In the last step, genome/cDNA databases are searched to identify genes that encode amino acid sequences determined by the previous step (*database search*).

After the first round of the above glyco-catch procedure, one may enter the second and third rounds for capturing other types of glycopeptides using different lectin columns (i.e. *lectin affinity-1′,1″*). For example, the flow-through fraction of concanavalin A (ConA) affinity chromatography in the first round is subjected to galectin affinity chromatography, and the flow-through fraction is then applied to peanut agglutinin (PNA)-agarose for the third round. By using serial lectin affinity chromatographies, different types of glycopeptides are efficiently recovered. This approach is in particular useful when less available biological sources are used as starting materials.

2.2. Core lectins used for glyco-catch procedures

Various lectins can be used to isolate glycoproteins and glycopeptides having distinct types of carbohydrate structures, e.g. galectins specific for LacNAc-containing glycans found exclusively in both *N*-glycans and *O*-glycans [26,28–30], ConA for oligomannosyl saccharides found in *N*-glycans [31,32], PNA specific for T-antigen found commonly in *O*-glycans [33], *Aleuria aurantia* lectin (AAL) showing broad specificity for L-Fuc-containing oligosaccharides [34], etc. Obviously, assessment of

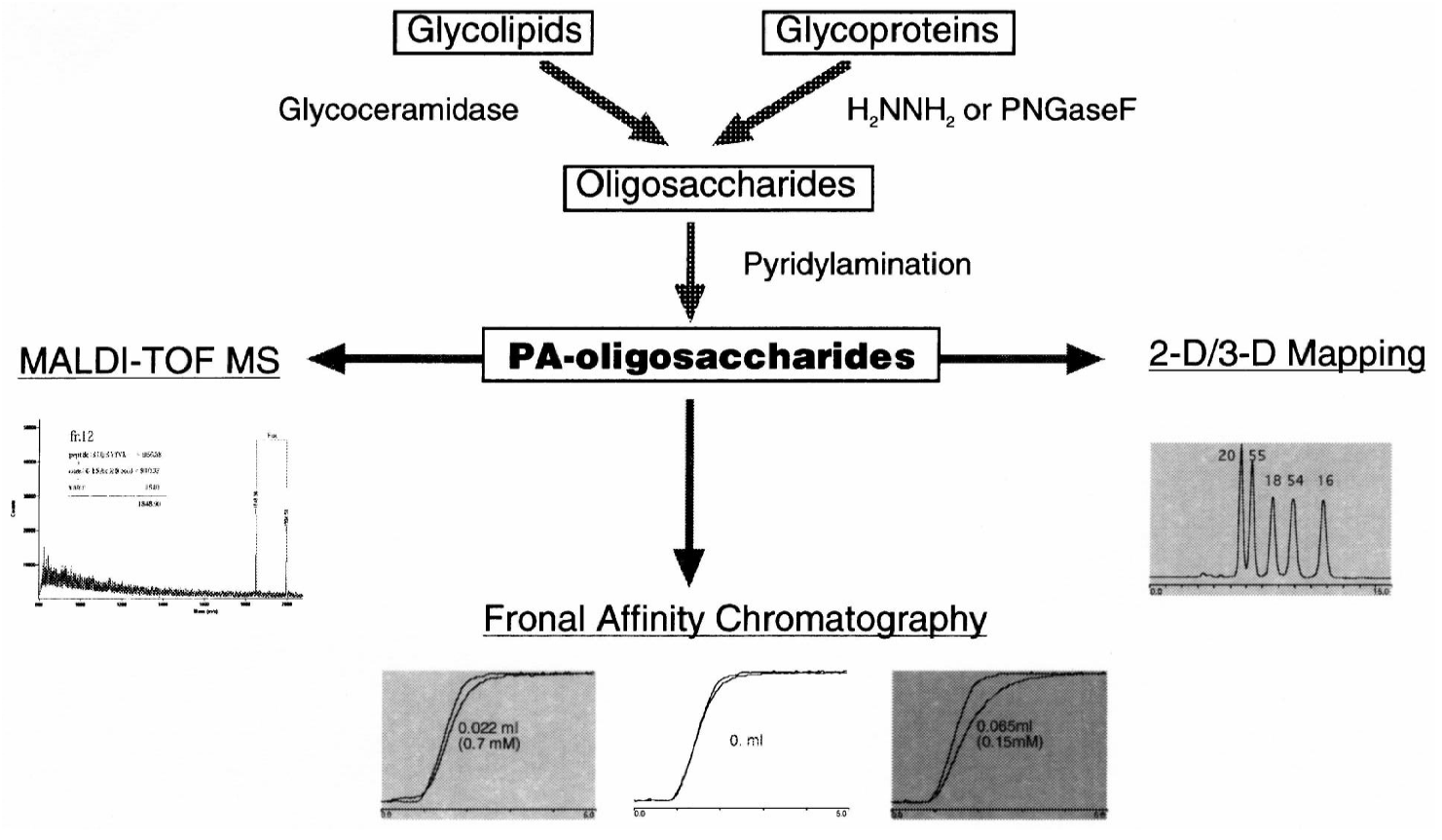


Fig. 2. Scheme of a glycan specification strategy. Glycans, derived by either chemical or enzymatic treatments of glycoproteins and glycolipids, are subjected to three characterization procedures based on distinct principles, all in the form of pyridylaminated derivatives: (1) MS analysis represented by MALDI–TOF-MS to obtain exact mass information, (2) 2-D/3-D mapping using normal-phase and reversed-phase HPLC (anion-exchange HPLC is added for 3-D mapping), and (3) reinforced frontal affinity chromatography to obtain affinity information (K_a) with respect to a set of lectins.

the lectins is most critical to achieve success in any glyco-catch procedure. In this context, it is essential to learn characteristics of some representative lectins, e.g. binding affinities, detailed sugar specificity, thermostability, metal dependency, multivalent features, etc. Here, some properties of core lectins used for the glycomic approaches undertaken in our laboratory are described. Basically, ConA and galectin LEC-6 (GaL6) bind specifically to high-mannose type [31,32] and complex type *N*-glycans [4,24,30]; respectively, whereas PNA binds with high affinity to the Galβ1-3GalNAc structure (T antigen) found in the core 1 structure of *O*-glycans [33]. Though their binding specificities are much more complex, these distinct types of lectins, when used, should provide us with a global view on the glycome of target cells and tissues. Detailed specificities determined for ConA by equilibration dialysis [3], and for GaL6 [24] and PNA [33] by frontal affinity chromatography in this work are schematically shown in Fig. 3.

However, one should have in mind that ConA and other polyvalent plant lectins (e.g. *Ricinus communis* agglutinin I) show considerably high affinity for a series of glycans, i.e. in the range of 10^6–10^8 M^{-1} in K_a [31,32]. On the other hand, affinities of galectins for lactosamine-containing saccharides are fairly weak, i.e. 10^3–10^6 M^{-1} in K_a [24,35]. Therefore, it is possible that glycopeptides having relatively low affinity may pass through some galectin columns. Moreover, most galectins cannot recognize any Le structure [4,35,36]. It is also possible that any substitution at the 6-OH group of the galactose moiety may abolish the binding. RCA-I is known to have high affinity for lactosamine (e.g. K_a for lacto-*N*-neotetraose, 10^5 M^{-1}). However, this lectin has only poor affinity for type 1 saccharides; and, like galectins, RCA-1 has only poor affinity for Le structures (J. Hirabayashi, T. Ueda, K. Kasai, unpublished result).

2.2.1. Concanavalin A (ConA)

ConA is a representative plant lectin that has high affinity for a series of high-mannose type *N*-glycans. It also binds to hybrid type *N*-glycans, and to a lesser extent to binantennary complex type *N*-glycans. Insight into the binding affinities of this lectin was obtained using a panel of oligosaccharides, and the results revealed that high-mannose type structures could be divided into three groups on the basis of affinity (Fig. 3, for structures, see Fig. 4): ConA binds most strongly to the glycans having the tetramannosyl core structure, ''Manα1-2Manα1-6(Manα1-3)Manβ'', with association constants (K_a) ranging from 3 to 5×10^7 M^{-1} [31,32]. This group

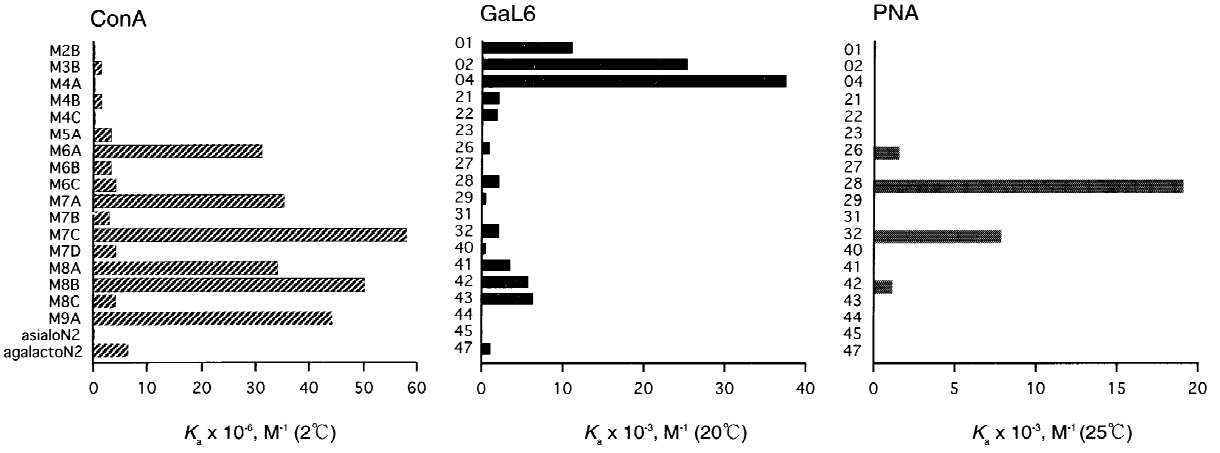


Fig. 3. Schematic representation of sugar-binding specificities of core lectins used in the present glyco-catch procedures. Association constants (K_a) to representative PA-oligosaccharides (for structures, see Figs. 4 and 5) are shown for comparison in bar graphs. Data on ConA were obtained by microdialysis at 2 °C [31], whereas data on GaL6 were originally obtained by previous frontal analysis at 23 °C [24], which were supplemented by data obtained at 20 °C in the present work. Data on PNA were obtained by the present frontal analysis at 25 °C.

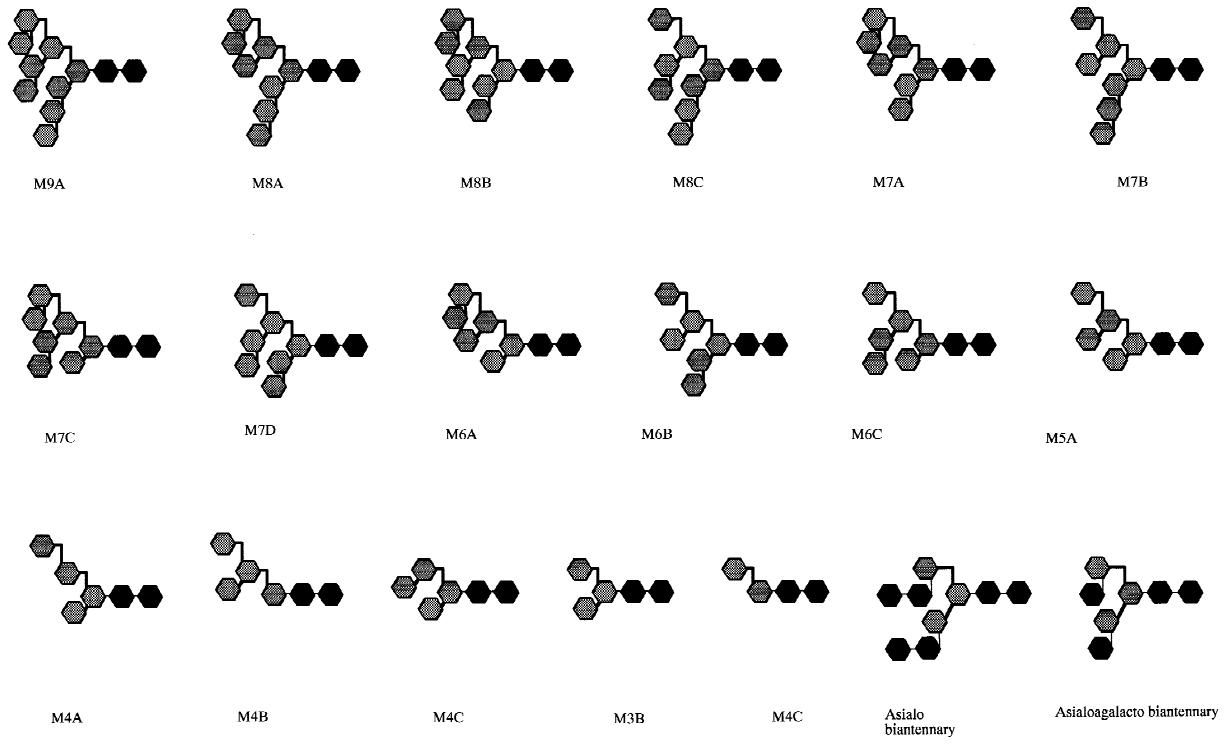


Fig. 4. Structures of *N*-glycans used for analysis with ConA in Fig. 3. A series of high-mannose type *N*-glycans and a few complex type *N*-glycans are schematically shown. ▨, ●, and ● represent component saccharides found in these structures; i.e. GlcNAc, Man, and Gal, respectively. Thick (▬) and thin (—) lines denote α- and β-linkages, respectively.

includes M6A, M7A, M7C, M8A, M8B and M9A. Apparently, the presence of non-reducing terminal Man, which is α1-2 linked to Man(α1-6), enhances the affinity by at least one order of magnitude relative to that of the trimannosyl core, ''Manα1-6(Manα1-3)Manβ''. Oligosaccharides containing this trimannosyl core constitute the second group (i.e. M3B, M4B, M5A, M6B, M6C, M7B, M7D and M8C), which shows moderate affinities in the range of $1–4\times10^{6}$ M^{-1}. Oligosaccharides lacking this trimannosyl core structure, e.g. M2B, or those having detrimental substitution at 3-OH or 6-OH groups of Man(α1-6), proved to have much weaker affinities [32] and represent the third group (i.e. M2B, M4A and M4C). X-Ray crystallography study of a complex between ConA and the trimannosyl structure (M3B in Fig. 4) clearly demonstrated the particular importance of this Man(α1-6) [37].

2.2.2. Galectin LEC-6 (GaL6)

Galectins are widely distributed, metal-independent, soluble lectins, whose binding specificity for β-galactosides is conserved among a broad range of multicellular organisms [28]. To the present, 12 mammalian galectins have been found and designated galectin-1 to -12. On the basis of their structural architecture, they are classified into three types, i.e. proto, chimera, and tandem-repeat types [29,38]. The presence of similar structural types of galectins have been annotated extensively in genome databases. Galectins in the nematode *C. elegans* are designated LEC-1 to -11, and LEC-1 was the first galectin investigated in invertebrates [29]. LEC-1 to -5 belong to the tandem-repeat type, having two homologous galectin-type carbohydrate-recognition domains (CRDs) that are characteristic of the galectin family; whereas LEC-6 belongs to the proto type, existing as a non-covalent dimer of a 16-kDa CRD. On the other hand, the remaining LEC-7 to -11 belong to a novel type [4]. Since galectin LEC-6 (designated GaL6 in this review) represents a major galectin in the nematode, its biochemical properties have been most extensively investigated [4,24,30].

The binding specificity of GaL6 has been analyzed

by means of frontal affinity chromatography (FAC, described in detail in Section 3.3), and the result is schematically shown in Fig. 3 (for structures, see Fig. 5; original data from Ref. [24]; additional data obtained in this study). GaL6 binds most strongly to branched complex-type *N*-glycans. As the number of branches increases, the binding affinity (K_a) for them also increases; i.e. **01** (NA2; $K_a=1.1\times10^4\ M^{-1}$)<**02** (NA3; $2.5\times10^4\ M^{-1}$)<**04** (NA4; $3.7\times10^4\ M^{-1}$). Apparently, this is the result of increased valency of the glycans. α2-6 sialylation of non-reducing terminal galactose abolishes the affinity; e.g. **23** (N2). Among glycolipid-derived glycans (numbers **26** to **47** in Fig. 3), GaL6 prefers type 1 structure (**42**, lacto-*N*-tetraose; $5.3\times10^3\ M^{-1}$) to type 2 structure (**41**, lacto-*N*-neotetraose: $3.1\times10^3\ M^{-1}$). Further, α1-2 fucosylation of the former saccharide, i.e. type 1 H antigen (**43**; $5.9\times10^3\ M^{-1}$), slightly enhances the affinity. However, α1-3GalNAc substitution at the non-reducing terminal galactose of **43**, resulting in A hexasaccharide (**47**; $6.2\times10^2\ M^{-1}$), considerably diminishes the binding power, though this oligosaccharide is strongly favored by mammalian galectin-3 [36] as well as by *C. elegans* galectin LEC-1 [35]. On the other hand, GaL6 cannot bind to Le antigens, e.g. **44** (Le^a) and **45** (Le^x) as is the cases for most other galectins. X-Ray crystallography studies on representative galectins, i.e. galectin-1 [39,40], galectin-2 [41] and galectin-3 [42] have proved the critical importance of 4-OH and 6-OH groups of Gal as well as the 3-OH group of GlcNAc for LacNAc (Galβ1-4GlcNAc) recognition.

2.2.3. Peanut agglutinin (PNA)

Peanut agglutinin (PNA) is known as a useful tool for histochemical studies to detect T antigen (Galβ1-3GalNAc), which is widely found in *O*-glycans of mucin-type proteins, such as leukosialin and glycophorin A [33]. This disaccharide unit forms the most fundamental ''core 1'' structure of *O*-glycans as a precursor of the ''core 2'' structure (Galβ1-3(GlcNAcβ1-6)GalNAc) and also as that of sialyl T antigen (NeuAcα2-6Galβ1-3GalNAc) [43]. Among glycolipid-derived oligosaccharides, asialoGM1 (Galβ1-3GalNAcβ1-4Galβ1-4Glc; No. **28** in Fig. 3) and GM1 (Galβ1-3GalNAcβ1-4(NeuAcα2-3)Galβ1-4Glc; No. **32**) contain this epitope structure. Frontal analysis was carried out to confirm this more quan-

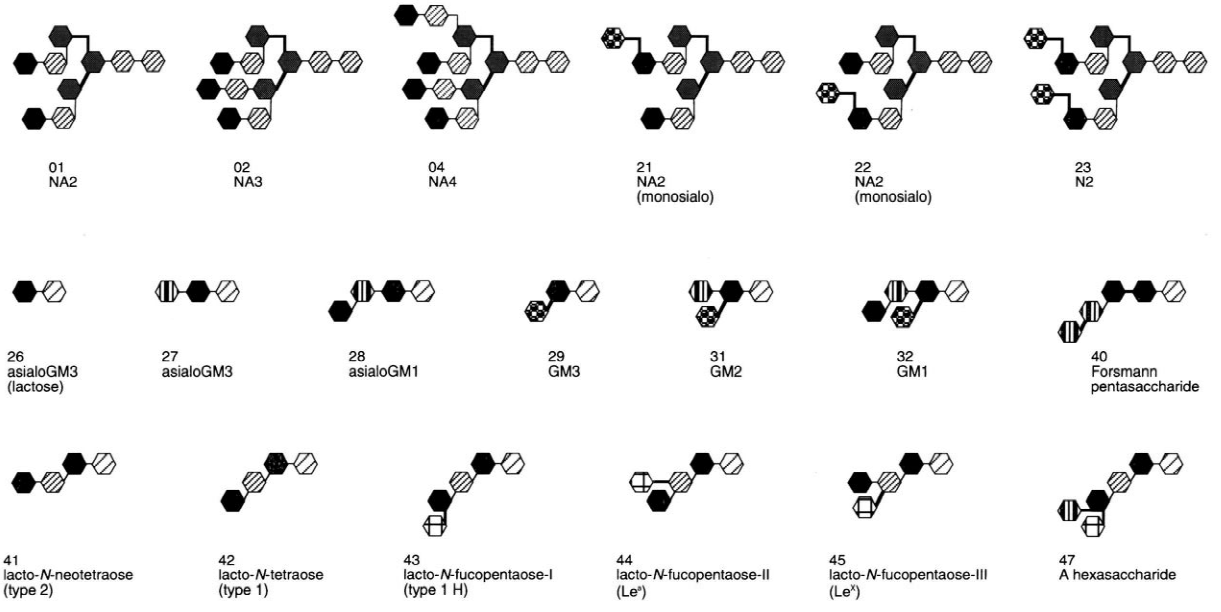


Fig. 5. Structures of galactose-containing glycans used for analyses with GaL6 and PNA in Fig. 3. A series of complex type *N*-glycans (**01**–**23**) and glycans originally derived from glycolipids (**26**–**47**) are schematically shown. ○, ◎, ●, ■, ◧, ◇ and ◈ represent component saccharides found in these structures, i.e. Glc, GlcNAc, Man, Gal, GalNAc, L-Fuc and *N*-acetylneuraminic acid, respectively. Thick (▬) and thin (—) lines denote α- and β-linkages, respectively, as in Fig. 4.

titatively. As a result, asialoGM1 (**28**; $K_a = 1.9 \times 10^4$ M^{-1} at 25 °C) showed the highest affinity for PNA, and GM1 (**32**; 7.7×10^3 M^{-1}) followed it. Lactose (**26**; 1.4×10^3 M^{-1}) and the type 1 saccharide lacto-*N*-tetraose (**42**; 1.0×10^3 M^{-1}) also showed significant but much weaker affinity. It is noteworthy that both galectins and PNA are classified as galactose-specific lectins simply on the basis of monosaccharide specificity. Nevertheless, they show completely distinct specificities; i.e. GaL6 binds most favorably to galactose-containing branched *N*-glycans (i.e. complex type *N*-glycans), whereas PNA selectively binds to the T-antigen commonly found in *O*-glycans.

For glycomic approaches, we chose the above three lectins as essential probes. Even if not perfect, this combination can provide us with a first glance of the glycome of target cells, tissues or organisms.

2.3. Model experiments

To confirm the validity of the thus developed glyco-catch method, we first attempted a few model experiments using representative glycoproteins having either high-mannose type or complex type *N*-glycans, i.e. chicken ovalbumin and bovine fetuin, respectively (for details, see Ref. [3]). Five milligrams of either of these glycoproteins was denatured prior to digestion with 1/100 (w/w) of *Achromobacter* protease I (commercially available as lysylendopeptidase from Wako Pure Chemicals, Tokyo, Japan). The use of this protease is strongly recommended for its rigorous specificity and high processivity [44]. By its high reliability, we can assume that captured glycopeptides are preceded by a lysine residue with only few exceptions, where derived peptides are N-terminal ones.

2.3.1. Ovalbumin

Ovalbumin has a single *N*-glycosylation site (amino acid position, 292), to which various high-mannose type or hybrid type *N*-glycans are attached [32,45]. Therefore, it is expected that ConA would bind tightly to glycopeptides having these glycans. The digest obtained by proteolysis of ovalbumin was applied to a ConA-column (column volume, 5 ml). The column was equilibrated with TBS (50 m*M* Tris–HCl, pH 7.5, 150 m*M* NaCl), and the bound glycopeptides were eluted with 0.1 *M* methyl-α-D-mannopyranoside dissolved in the same buffer. Eluted fractions were subjected to reversed-phase chromatography (Fig. 6, *left*), followed by sequencer analysis (for details, see Ref. [3]). The ovalbumin digest gave a single major peak with the retention time of 20 min by reversed-phase chromatography on a TMS-column. The sequence completely corresponded to a peptide 291–322, which contained a previously identified *N*-glycosylation site (Asn^{292}–Leu–Thr) [44]. In this case, Asn^{292} was not detectable as an intact phenylthiohydantoin-amino acid, even though the flanking positions were successfully determined in the sequence analysis.

2.3.2. Fetuin

Fetuin has three *N*-glycosylation sites (amino acid positions 99, 156, and 176), to which mainly triantennary complex type *N*-glycans are attached [46,47]. Therefore, the *C. elegans* galectin GaL6 should bind to these glycans, unless the 6-OH group of the non-reducing terminal galactose is masked by sialic acid. For this reason, fetuin was desialylated by acid treatment (80 °C, 1 h) prior to digestion with lysylendopeptidase. The GaL6-column was equilibrated with MEPBS (4 m*M* β-mercaptoethanol, 2 m*M* EDTA, 20 m*M* Na-phosphate, pH 7.2, 150 m*M* NaCl). After application of the digest, the column was extensively washed with the same buffer, and the adsorbed glycopeptides were eluted with 0.1 *M* lactose in MEPBS. The eluted fraction was subjected to HPLC separation as described above. The main peak, which eluted with a retention time of 28 min, was found to be a mixture of equal amounts of two peptides; i.e. 145–211 (peptide 1) and 219–225 (peptide 2; Fig. 6, *right*). The longer peptide was shown to have 2 *N*-glycosylation sites (Asn^{156}–Asp–Ser, and Asn^{176}–Gly–Ser; *underlines* denote the glycosylation sites), whereas the shorter peptide had no such site. This implies that these two peptides were linked by a disulfide bridge [45]. In the sequence analysis, neither Asn (corresponding to the 12th and 32nd cycles) was detected. This observation implies that Asn^{156} and Asn^{176} were glycosylated, as earlier reported [45]. Other smaller peaks (having reduced retention times) also contained part of the sequence 145–211, but not any of the 219–225 one. Therefore, these smaller peaks are attributed to only the peptide 1 as a result of cleavage of a disulfide bridge (Cys^{207}–Cys^{219}). To circumvent such a com-

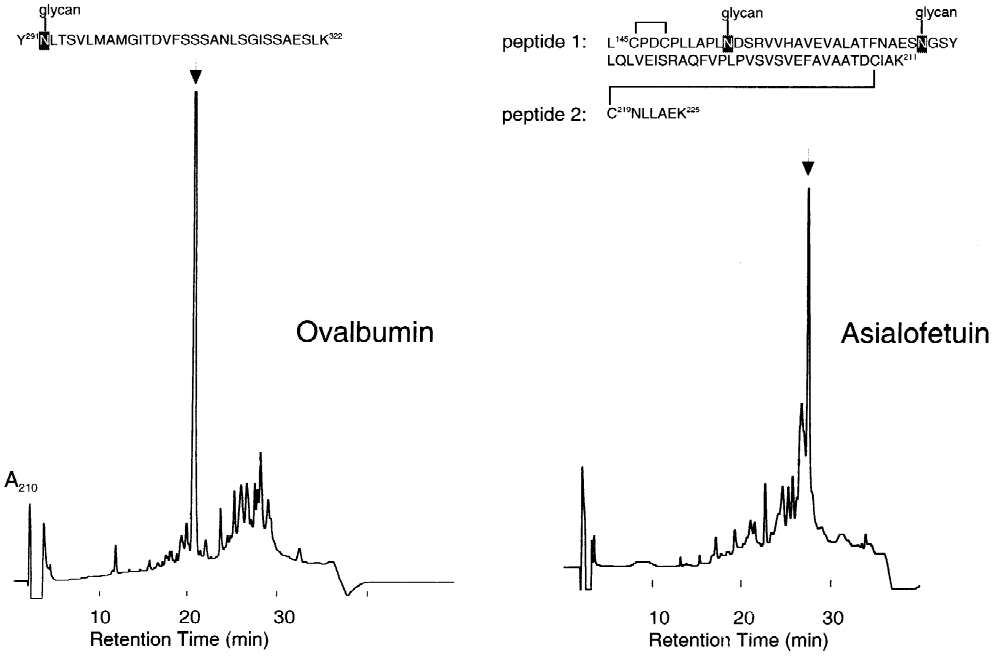


Fig. 6. Model experiments to demonstrate the validity of the proposed glyco-catch method. Ovalbumin containing high-mannose type *N*-glycans and asialofetuin containing complex-type *N*-glycans were digested with *Achromobacter* protease I (lysyslendopeptidase) in the presence of 4 *M* urea, and the target glycopeptides were captured by ConA-agarose (*left*) and GaL6-agarose chromatography (*right*), respectively. Derived glycopeptides were separated by reversed-phase chromatography on a Develosyl-TMS column (4.6×150 mm, Nomura Chemical, Tokyo, Japan) by a linear gradient of acetonitrile in 0.1% trifluoroacetic acid. Chromatography was carried out at a flow-rate of 1 ml/min at room temperature (22–23 °C). Peptides were detected by their absorbance at 210 nm (approximate full scale 1.0). The figure is reformed from Fig. 3 in Ref. [3].

plicated situation, we added a reduction process prior to a glyco-catch procedure when using a *C. elegans* soluble extract (described below). On the other hand, the remaining glycopeptide containing Asn^{99} was not recovered by this experiment.

Despite the report of various glycoforms in ovalbumin [45] and asialofetuin [47], both of these glycopeptides appear to be eluted as a single peak. Therefore, glycan moieties likely contribute to a lesser extent to peptide retardation in reversed-phase chromatography. In both cases of ovalbumin and asialofetuin, all of the identified peptides were preceded by a lysine, thus confirming the rigorous specificity of *Achromobacter* protease I.

2.4. Application to C. elegans soluble glycoproteins

As a more practical approach, we applied the glyco-catch method to a crude extract derived from *C. elegans* (data submitted). A soluble extract from 10 g of the worms was first applied to a ConA-agarose column (bed volume, 20 ml). Its flow-through fractions were pooled and then applied to a GaL6-agarose column (15 ml). Adsorbed proteins were eluted with either α-methyl-mannoside (ConA) or lactose (GaL6), respectively. As a result, 23.1 and 3.9 mg of glycoproteins, respectively, were obtained. On the other hand, when the extract was applied first to the GaL6 column, and then its flow-through fractions were applied to the ConA column, the yields became 5.5 mg (GaL6) and 20.2 mg (ConA), respectively (data not shown). The slight increase in the yield of the GaL6-adsorbed fraction might be attributed to the presence of glycoproteins having hybrid type *N*-glycans or of those having both high-mannose and complex-type ones, which would be captured by ConA. In any case, most (>75%) of the soluble glycoproteins present in mixed stages of *C. elegans* are considered to have high-mannose type

N-glycans. Obviously, compositions of the proteins captured by ConA and GaL6 differed significantly (Fig. 7). However, it should be noted that the eluted proteins were not necessarily glycoproteins specifically recognized by these lectins, because of possibility that non-glycosylated proteins were bound to target glycoproteins via protein–protein interaction or non-target glycoproteins were bound to target glycoproteins via cross-linking lectins. Glycoproteins obtained by the former procedure (i.e. ConA→GaL6) were digested with *Achromobacter* protease I, and the resultant glycopeptides were recovered by using the same lectin columns to increase the specificity. The captured glycopeptides were fractionated by reversed-phase chromatography

Fig. 7. Analysis of affinity-purified glycoproteins by SDS–polyacrylamide gel electrophoresis. Proteins adsorbed on ConA, GaL6, and PNA columns were eluted with 0.1 M α-methyl-mannoside (for ConA) or 0.1 M lactose (for GaL6 and PNA), respectively, and were subjected to electrophoresis on a 14% gel. Protein was stained with silver. As regards the PNA fraction, results from the EtOH precipitate (ppt) and lysylendopeptidase digest (Lys) are shown as well as that from the lactose eluate (lac).

as described above (Fig. 8). All fractions between 10 and 50 in each chromatography were subjected to sequence analysis. Corresponding nucleotide sequences were searched for in the *C. elegans* genome database by using TBLAST-N [48] or recently developed SQMATCH programs (http://ftp2.ddbj.nig.ac.jp:8080/sqmatch/ja.html).

2.4.1. Glycoproteins captured by ConA–agarose

By the proposed glyco-catch procedure using a ConA-agarose column, a number of genes have successfully been identified (details will be published elsewhere). In summary: (1) 32 genes were assigned for 44 peptides (designated C1–44 in Fig. 8). (2) These peptides were preceded by a lysine with three exceptions (C3, C19, and C24). All of them was preceded by putative signal sequences ending with Ala. (3) The analyzed peptides had a single N-glycosylation site (i.e. Asn–X–Ser/Thr, where X is any amino acid except Pro) with two exceptions, i.e. peptides C16 and C37. C16 had two potential N-glycosylation sites and peptide C39 lacked such a site. Inclusion of the latter peptide might be attributed to either accidental contamination or miss-prediction of an open reading frame of the gene. Alternatively, the peptide may not have been really glycosylated but just had some affinity for ConA. (4) Among the 32 assigned genes, 24 (78%) encoded secreted proteins, and four encoded membranous proteins, i.e. one type-I, two type-II and one multiple trans-membrane proteins. These proteins may have been obtained as a result of either partial proteolysis or alternative splicing. The remaining four (F20D6.4, C18H7.1, T06D8.1 and K07E12.1) had no predicted signal sequence, and thus were classified as cytoplasmic proteins. However, this conclusion is apparently contrary to the present result, because they were specifically captured by the ConA-column, and thus must have been secreted after glycosylation in the endoplasmic reticulum and Golgi apparatus. As far as F20D6.4, C18H7.1 and K07E12.1 are concerned, they are not likely to be cytoplasmic proteins as predicted, but extracellular proteins, because they have homologues in other species that function extracellularly, i.e. Kunitz family protease inhibitor (F20D6.4), von Willebrand factor A domain (C18H7.1), and a cell adhesion molecule (K07E12.1). Whereas 26 genes including these three

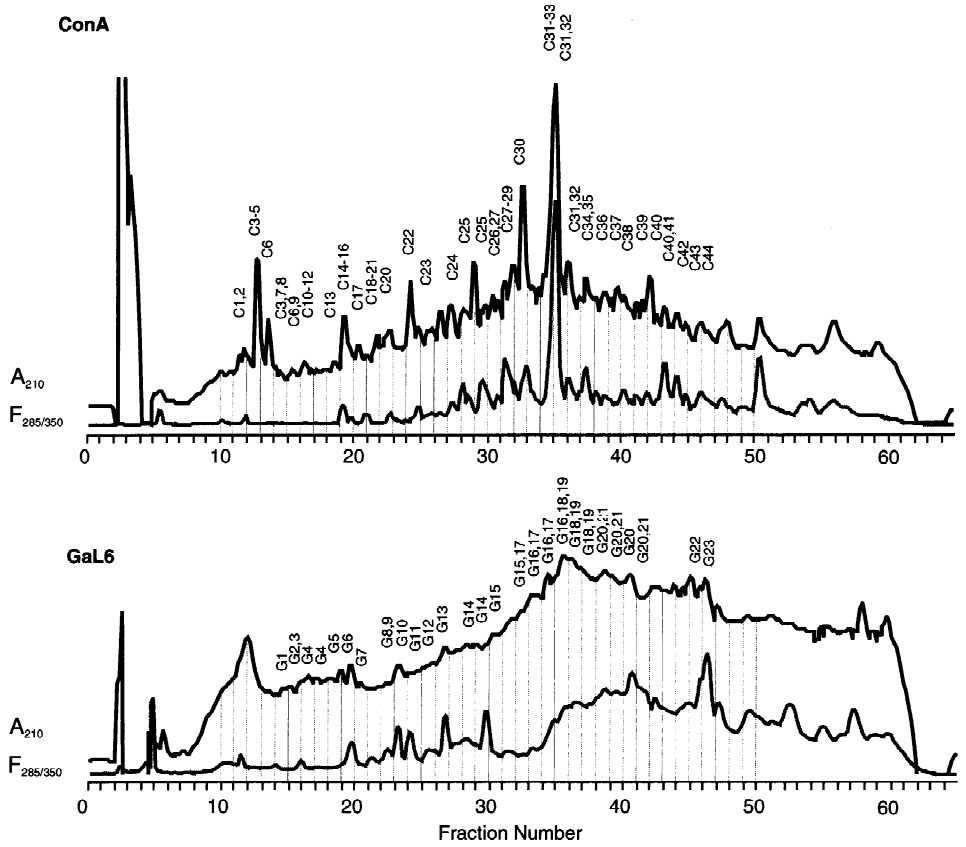


Fig. 8. Separation of glycopeptides captured by ConA (*top*) and GaL6 (*bottom*). Reversed-phase chromatography was carried out on a Develosyl-TMS column as described in Fig. 6. An aliquot of each fraction (1 ml) was subjected to sequencer analysis, and peptides, of which corresponding genes were successfully assigned, are designated C1–C44 (ConA) and G1–G22 (GaL6).

encoded proteins showing homologues in other species having extracellular functions, the remaining six had no apparent homology to any registered proteins.

Among the glycopeptides specified by identified genes, glycopeptides C3 (1208 pmol) encoded by Y5F2A.1 (transthyretin) and C31 (644 pmol) encoded by C42D8.2 (*vit-2*, vitellogenin) were recovered in the highest yields (Table 3). On the other hand, cosmids C25A1.8 (C-type lectin) and K07H8.6 (*vit-6*, vitellogenin) assigned by analysis of peptides C26 and C37, respectively, represent abundantly expressed genes, because they have many registered ESTs; i.e. 113 (C25A1.8) and 84 (K07H8.6) ESTs. These results imply that peptide recoveries and EST numbers are not necessarily in a direct correlation. A possible reason for this discrepancy may be partly attributed to diversity in glycan structures, because ConA recognizes diverse glycans with different affinities, as was discussed in Section 2.2.1. Among the genes identified using this lectin, C37C3.6 is unique in encoding a protein having a highly glycosylated state. As many as seven glycopeptides were obtained from this protein consisting of 1558 amino acids. Apparently, C37C3.6 specifies a heavily glycosylated protein with high-mannose type glycans. The described glycopeptides are listed in Table 2, along with some features of their genes.

2.4.2. Glycoproteins captured by GaL6-agarose

Compared with the above results obtained with ConA-agarose, relatively few genes have been identified by analysis of glycopeptides adsorbed on a GaL6-agarose. This is mainly due to a lower level of

Table 3
Features of representative glycopeptides captured by ConA-agarose and assigned genes

Fr.	AA sequence*	Yield (pmol)	Cosmid (Chromosome)	EST	ORF (aa)	Signal regions	Homologues in other species
C1	NGTGCGENCL TTK	68	C37C3.6 (V)	22	1558	1-28	Kunitz type trypsin inhibitor
C3	RLQNVTVK	1208	Y5F2A.1 (IV)	3	132	1-19	Transthyretin
C6	IAACNQTQES GTVCGAGYK	188	C37C3.6 (V)	22	1558	1-28	Kunitz type trypsin inhibitor
C15	TRRVICAIHQ NGGLEVVDEG HCQAEKPEGK	140	C37C3.6 (V)	22	1558	1-28	Kunitz type trypsin inhibitor
C23	NGSDHFYLNG NGLIQVEK	144	C37C3.6 (V)	22	1558	1-28	Kunitz type trypsin inhibitor
C26	DLGTTAFEDI SFPARAPPAF VNQTTEK	56	C25A1.8 (I)	113	233	1-25	C-type lectin
C27	TIAFGPHYSG CERSSFPCEL SDFGCCPDGE TAALGKNGTG .....LTTK	168	C37C3.6 (V)	22	1558	1-28	Kunitz type trypsin inhibitor
C28	SCETVDCEAE WFTGDWESCS STCGDQGQQY RVVYC..... NCTVERPPVK	56	C37C3.6 (V)	22	1558	1-28	Kunitz type trypsin inhibitor
C31	TYNYWTVSSR PENNENDRVV VQLTVEPPMSR QYVNITMQSP MERIELK	644	C42D8.2 (X)	61	1613	1-20	Vitellogenin
C32	LLAADACPAD EQEKFDTERT CNLGPCEGLT FV....RCND TEETREVTCK	422	C37C3.6 (V)	22	1558	1-28	Kunitz type trypsin inhibitor
C36	NMIRATINVE PRQRLTVNMT IETPMETTVL ERVELPFRLP ....SRYEQK	36	K07H8.6 (IV)	84	1651	1-19	Vitellogenin
C37	MSSTVLNYNI TGTSSSFLIN SVELRSQHLF APISEK	64	K07H8.6 (IV)	84	1651	1-19	Vitellogenin

Underlines denote regions for which sequence analysis was carried out. Asparagine residues that were confirmed to be glycosylated by the absence of PTH-Asp in corresponding cycles are shown as N, whereas those remaining as potential *N*-glycosylation sites, because they were out of the range of sequence analysis, are shown as N

biosynthesis of glycoproteins containing complex type *N*-glycans in *C. elegans*, because at most 5 mg of GaL6-binding glycoproteins were obtained from 10 g of the worm, whereas >20 mg of ConA-binding soluble glycoproteins were obtained.

As a result of sequence analysis of these glycopeptides: (1) 16 genes were assigned to 23 peptides (G1–G23 in Fig. 8). (2) These peptides were preceded by a lysine with a single exception (G7), which was preceded by a putative signal sequence ending with Ala, as in C3, C19, and C24. (3) Among the 23 analyzed peptides, 17 had a single *N*-glycosylation site, four (G16, G17, G19 and G20), had multiple sites, and the remaining one (G18) lacked such a site. However, G18 seems to have been covalently linked to glycopeptide G19 by disulfide bond(s), because both G18 and G19 were encoded by the same gene, F57F4.3, and had multiple cysteine residues. Moreover, both peptides were eluted in the same fractions in similar yields by reversed-phase chromatography (Fig. 8). (4) Among the 16 assigned genes, nine encoded secreted proteins, whereas four encoded membranous proteins consisting of two type-II and two multi-membranous proteins. Capturing of the peptides derived from these membranous proteins may have been the result of either partial proteolysis or alternative splicing as described above. The remaining three genes (Y49E10.18, Y4C6B.6, T28F3.8) encoded no signal sequence, though their predicted functions are likely to be extracellular ones, e.g. lipase (Y49E10.18) and glucosidase (Y4C6B.6). 5) Among the 16 assigned genes, 11 encoded proteins showing homologues in other species, whereas the remaining five had no apparent homology to any of the presently known proteins.

Among the identified genes, F57F4.3 (EGF-like repeat) is unique for the same reasons given for C37C3.6. The gene F57F4.3 has 32 ESTs in the *C. elegans* genome database, and its predicted protein consists of 2153 amino acids, and has significant homology to proteins with EGF-like repeat. The protein is also predicted to have a GPI-anchoring site in its C-terminal region (Ser^{2061}) (http://www.proteome.com/databases/). More critically, the protein has as many as 36 potential *N*-glycosylation sites. Among them, we identified four sites by direct analysis of G2, G16, G17, and G21, whereas 10 sites found in G16, G17, G19, and G20 remain as potential sites, though they are strong candidates for glycosylation. On the other hand, the asparagine residue identified at the 12th position of glycopeptide G20 proved not to be glycosylated. Analyzed peptides encoded by F57F4.3 are shown in Table 4, along with some gene features.

2.4.3. Glycoproteins captured by PNA-agarose

This type of glycan is usually termed a ''mucin type'', since it is widely found in mucus tissues, such as gastrointestinal tissues and submaxillary gland. The glyco-catch method has proved to be successful in identification of *N*-glycosylated proteins as described above. However, its direct application to *O*-glycosylated proteins would be expected to meet

Table 4
Glycopeptides encoded by F57F4.3 captured by GaL6-agarose

Fr.	AA sequence	Yield (pmol)	Cosmid (Chromosome)	EST (aa)	ORF regions	Signal	Homologue in other species
G2	VHNNCTSPEA GVTACCCDSD ACLDPNRGK	159	F57F4.3 (V)	32	2153	1-19	EGF-like repeat
G16	NATVYYCAPV GMCRYPGLGF GGSYGNCGQI PGADPQVTGC ....NRVPGK	14					
G17	GDCVAVNLNT TYNGVATTAS LYTCDPSYIC RMMNSTNRCH ....STPPRK	34					
G18	DTCRPLWSDR EVTACCCNNA DNCNLK	23					
G19	DPNVKPGPAV LPDFPTACYQ GLLVNNQTYG APLTLQGCYG ....YPGPAK	7					
G20	CASVNARIAN DNVTLFACVP HSLCRSLELY DSCARMEPYY ....VCRNLK	25					
G21	INTSMPVTNF RDYPIACFSG LYVNNMPISI AGWQACK	14					

with technical difficulties, because *O*-glycosylation often occurs at multiple sites of clustered regions that are rich in Ser and/or Thr in an extremely heterogeneous manner. Even if not glycosylated, Ser and Thr represent the most difficult amino acids to determine by sequencer analysis due to their poor recoveries. Therefore, resultant *O*-glycosylated peptides captured by the procedure using PNA-agarose will not be easily identified by the proposed strategy. Despite such difficulties, we have tried the glyco-catch procedure and succeeded in identification of a gene (K07E12.1) by querying the following sequence: VIP(T)DE(S)GSVVYPITK (*underline* denotes amino acids detected by sequencer analysis). The peptide was preceded by a lysine, and the fourth position (Thr) was not detectable at all, and thus, is most likely to be an *O*-glycosylation site. This is ensured by the fact that both flanking Pro and Asp could be identified unambiguously. In addition, occupation at −1 position by Pro relative to the assumed *O*-glycosylation site (Thr) agrees with the empirical rule of a preponderance of proline residues at ''−1'' and/or ''+3'' positions. K07E12.1 is predicted to encode an extremely large extracellular type protein consisting of 13 055 amino acids, but this prediction has not been confirmed experimentally. Importantly, the predicted gene product shows 21% homology to a major human mucin, MUC-2 [49].

However, we have not succeeded in identification of genes by direct analysis of other glycopeptides captured by the PNA-agarose for the reasons given above. A possible solution for this problem is *O*-deglycosylation before HPLC separation. Since *O*-glycosidase treatment can fully liberate *O*-glycans from Ser/Thr residues of glycopeptides, thus deglycosylated peptides would be expected to be easily identified by either protein sequencer or concurrent LC–MS identification procedures (Kaji et al., this issue). In the latter procedure, information on *O*-glycosylation sites will be derived by skilled incorporation of $H_2{}^{18}O$ upon hydrolysis, as originally described by Gonzalez et al. for identification of *N*-glycosylation sites [50].

3. Characterization of glycans

''How to describe glycans'' is a dominant issue of structural glycomics, because conventional procedures to define covalent glycan structures are not very sensitive and are too laborious. Therefore, development of a novel procedure to define glycan structures more efficiently is keenly awaited. Straightforward interpretation of the ''glycome project'' is to determine thousands of thousands of covalent structures of all glycans produced in an organism. However, for this achievement we must realize that glycans are synthesized in a considerably different manner than genes and proteins as described in Section 1.

Here, we propose a novel strategy to specify glycans in the context of structural glycomics, where the following three criteria are defined as essential information:

1. Physical mass information derived by mass spectrometry (MS)
2. Chemical structural information derived by 2-D/3-D mapping
3. Bio-affinity information obtained by frontal affinity chromatography (FAC)

First of all, it should be emphasized that all of these procedures are based on completely different principles, but can be well performed on pyridylaminated (PA) oligosaccharides [27]. MS analysis gives exact molecular mass, whereas 2-D/3-D mapping extracts refined chemical information, i.e. molecular size and hydrophobicity in terms of glucose units. On the other hand, FAC analysis provides us with reliable affinity information, i.e. association constant K_a (or dissociation constants K_d) between each glycan and a certain set of lectins, such as ConA, GaL6 and PNA, for which detailed specificities are well known. The proposed strategy to define glycans is still preliminary, but the essential scheme can be sketched as in Fig. 2.

3.1. TOF-MS analysis

There is no need for explaining the necessity for MS. Molecular masses are the most fundamental requisites to define chemical structures of glycans. From a technical viewpoint, intact glycans are only poor analytes for any MS procedure because of their low volatility. However, derivatization with appropriate hydrophobic reagents greatly increases the sensitivity. Various methodologies for ionization and detection systems have been developed for proteomics, and they will be further improved in the future. Hence, MS and MS–MS technologies are essential items for glycomics, too.

3.2. 2-D/3-D mapping

As regards 2-D/3-D mapping, Takahashi and coworkers have developed a refined system using anion-exchange, normal-phase and reversed-phase chromatography of PA-oligosaccharides to not only separate but also identify diverse glycans [22,23]. Up to the present, >400 PA-glycans have been mapped with coordinated Glu units. 2-D plots of >200 PA-*N*-glycans (data originally from http://www.gak.co.jp/ECD/Hpg_eng/hpg_eng.htm) are shown in Fig. 9. Identification procedures by HPLC are conventionally carried out by a linear gradient elution. However, such a system is rather time-consuming (>80 min per analysis) and not very reproducible. Therefore, we introduced isocratic elution for each type of glycan (i.e. high-mannose type, complex type and glycolipid-derived glycans). The isocratic system is rapid (<20 min) and reproducible, and thus, will be favored for future studies in glycomics.

3.3. FAC (frontal affinity chromatography)

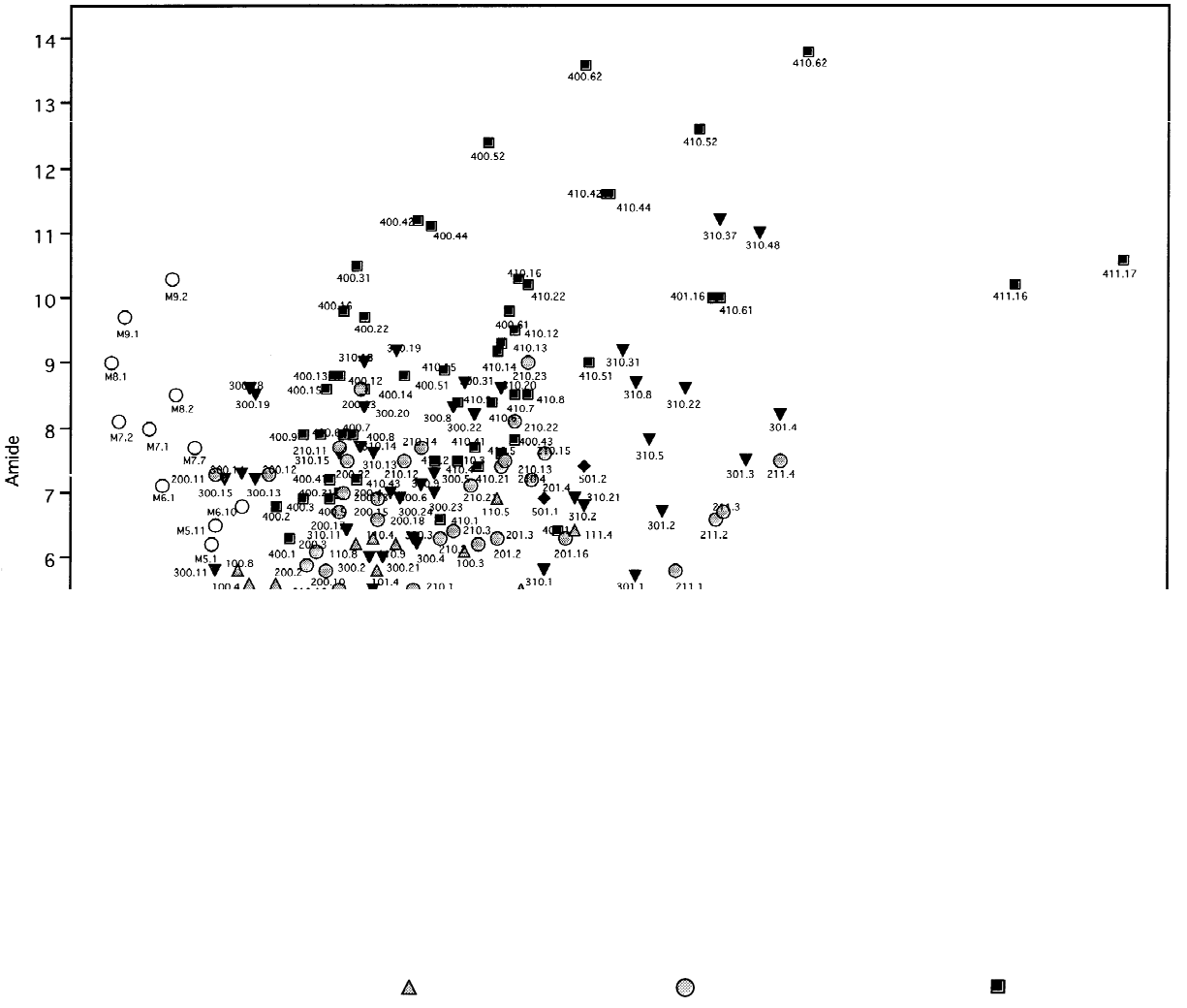
The most distinguished feature of the proposed glycome project may be represented by the use of FAC. FAC is a quantitative method developed a quarter century ago by Kasai and Ishii (for review, see Ref. [51]), which can accurately determine affinity constants (K_a) between various biomolecules, such as enzyme–substrate analogues (e.g. inhibitors) [52,53] and lectin-oligosaccharides [32,35,54,55]. Other methods to determine K_a include classic equilibrium dialysis, and more recently, a biosensor technique based on surface plasmon resonance [56,57]. However, most of these methods have some difficulties in either accuracy, simpleness or economy. In this context, recently reinforced FAC has greater merits, in the first place, in terms of its clarity and accuracy [24,25,35]. The system is also desirable from both economical and operational viewpoints.

3.3.1. Principle

Frontal affinity chromatography is a quantitative method to determine dissociation constants K_d-values (or association constants K_a-values) between two biomolecules (A and B) by means of an affinity chromatography [51]. As a standard procedure, an excess volume of an analyte A (e.g. PA-oligosaccharide) is applied at an initial concentration of $[A]_0$ (M) to a column, on which affinity ligand B (e.g. GaL6) is immobilized (Fig. 10a). In comparison with a negative control (e.g. Rha having no affinity for GaL6), the difference in volume of the elution fronts, i.e. $V_f - V_0$ (ml) reflects the affinity between A (PA-oligosaccharide) and B (GaL6) as defined by Eq. (1) (Fig. 10b), where B_t is the effective ligand content (mol) for the packed column. The basic equation of FAC (Eq. (1)) is given as follows [51]:

$$K_d = [A][B]/[AB] = B_t/(V_f - V_0) - [A]_0 \quad (1)$$

For the determination of K_d (M) and B_t (mol) either the Lineweber–Burk type (Eq. (2)) or the Woolf–Hofstee type plots (Eq. (3)), both of which


Amide
14
13
12
11
10
9
8
7
6

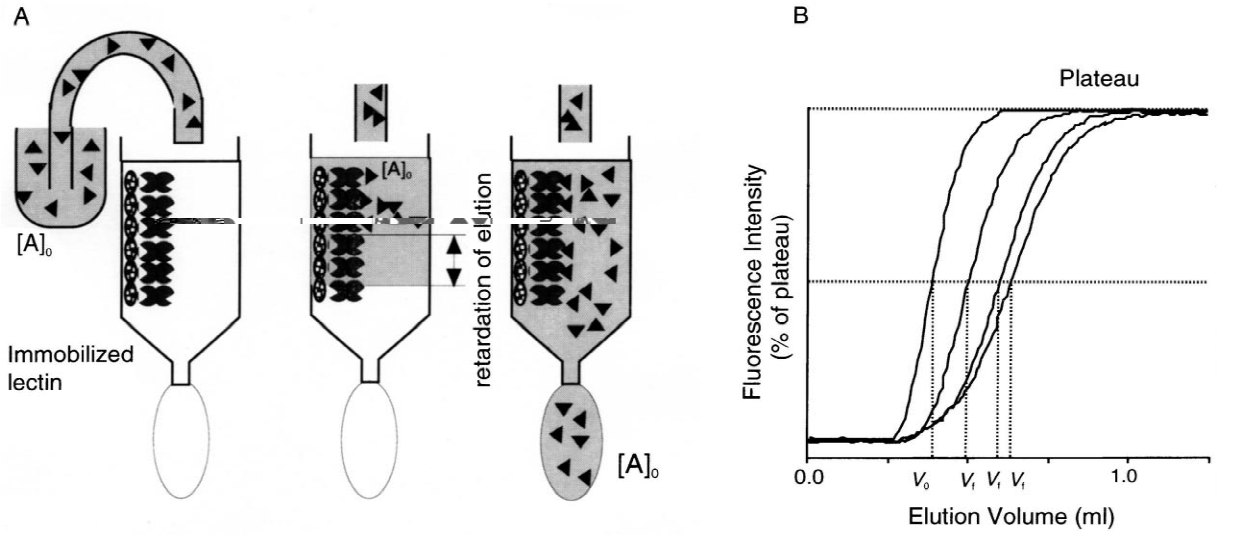


Fig. 10. (A) The principle and operation of frontal affinity chromatography (FAC). In FAC, an excess volume of an analyte, A, is continuously applied at a concentration of $[A]_0$ (initial concentration) to a column, on which affinity ligand B is immobilized. If A has some affinity for B, the elution front of A is retarded compared with that of a sample with no affinity for B. (B) Examples of elution profiles of PA-oligosaccharides, where V_0 is obtained by elution of a control saccharide, Rha. The difference in volumes of their elution fronts, i.e. $V_f - V_0$, reflects the affinity between A and B by the basic equation of FAC (Eq. (1), see text). The figure is reformed from Fig. 2a in Ref. [3].

pretation of chromatograms obtained by FAC is extremely easy.

$$K_d = B_t/(V_f - V_0) \quad \text{if } [A]_0 \ll K_d \tag{4}$$

$$K_a = (V_f - V_0)/B_t \tag{5}$$

It should be emphasized that FAC maximally exerts its potential for analysis of weak interactions. These include most sugar–lectin interactions (10^3–10^6 M^{-1} in K_a), which is in general weaker than antigen–antibody interaction (10^6–10^{10} M^{-1} in K_a). Despite these merits, previous FAC procedures required a relatively long time for operation (e.g. >3 h per analysis) and large amounts of samples for analysis, mainly because conventional open columns were used. Recently, these shortages were dramatically improved in two laboratories [24,58]. Hirabayashi et al. [24] introduced various merits of HPLC to reinforce a system for FAC: They used a miniature column (10×4 mm; bed volume, 0.126 ml) packed with lectin-immobilized resins, to which fluorescently-labeled oligosaccharides are applied at a constant flow-rate (usually 0.25 ml/min), and at a constant concentration (10 n*M* in case of PA-oligosaccharides) via a relatively large (2 ml) sample loop (Fig. 11). For determination of V_f, a data processing program using Microsoft Excel has also been developed [25].

3.3.2. Application to glycomics

As described above, FAC is a simple method to determine K_a-values (K_d-values) with high accuracy and reproducibility for pairs of lectins and glycans. Thus, derived lectin-affinity information should greatly contribute to define glycan structures. The procedure can be applied to glycans derived from either glycoproteins (e.g. by hydrazinolysis or peptide-*N*-glycanase F treatment) or glycolipids (e.g. by glycoceramidase treatment; see Fig. 2). Even if covalent structures are not determined, proteome/glycome databases should work with thus obtained sets of valuable information; i.e. (1) cosmid ID, (2) glycosylation sites, (3) lectins used for glyco-catch procedures, (4) yields of captured glycopeptides, (5) sources of glycoproteins, (6) mass values of liberated glycans, (7) behaviors in 2-D/3-D mapping in terms

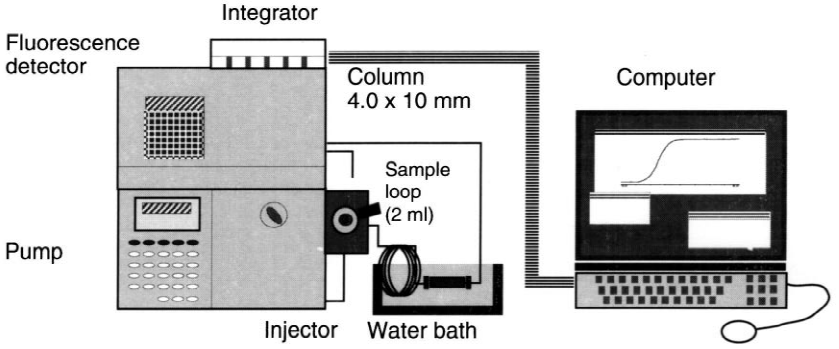


Fig. 11. The system for reinforced FAC [3,24]. A conventional HPLC system is used with a relatively large (2-ml) sample loop and a relatively small column (4×10 mm) so that an excess volume of an analyte is continuously applied to a column at a constant concentration (10 n*M* in case of PA-oligosaccharides) and at a constant flow-rate (0.25 ml/min routinely). For sensitive analysis, a fluorescence detector is used. For accurate determination of the volume (ml) of the elution front of each sample, a data processing program using Excel has also been developed [25]. The figure is originally from Fig. 2b in Ref. [3].

of glucose units, and (8) K_a values for a set of lectins.

If one assumes that lectins are ''decipherers of glycans'' [26], FAC may be used to define each glycan structure in a more innovating manner. If the best combination of multiple lectins is refined to probe various elements of glycan structures (i.e. modules), e.g. (1) Gal-containing or not? (2) if yes, βGal or αGal? (3) if βGal, Galβ1-3GalNAc (T antigen), Galβ1-3GlcNAc (type 1) or Galβ1-4GlcNAc (type 2)? (4) branching or not? (5) if yes, the number of branches? (6) the type of branching? (7) extension by poly-*N*-acetyllactosamine? (8) terminal modification with Fuc, Sia, Gal or GalNAc?, such a multiple mapping system on the basis of ''pattern recognition'' should work for future glycomics.

4. Technical problems

From a technical viewpoint, there are some problems in the proposed strategy for glycoprotein-targeted glycomics: (1) limitation to gene identification efficiency by the present system using a protein sequencer. In this regard, a recent proteomic trend using either LC–MS–MS [59] or CE–MS–MS [60,61] will be most promising. In this context, however, the present gene-identification program developed for proteomics should be improved for glycopeptides, because molecular masses of glycan moieties are not easily predicted at the in silico level. (2) Imperfectness to recover glycopeptides. As described, the use of lysylendopeptidase is strongly recommended from a practical viewpoint, but it does not mean that the enzyme works perfectly in all aspects. In fact, generated glycopeptides which are too short or too large are difficult to recover in reversed-phase chromatography. In case of very short glycopeptides (i.e. <6 amino acids), it is impossible to identify a target gene in the genome database. Moreover, sequence analysis by a protein sequencer can usually be carried out not more than 20 cycles. Therefore, in case of too large glycopeptides (i.e. >20 amino acids), even if their analysis leads to successful identification of target genes, limited information is obtained as regards glycosylation sites (as described in Section 2.4.2). If only a single potential site is included in the captured glycopeptide, one may conclude that the potential site is a ''logical'' glycosylation site. However, if the captured glycopeptides contain multiple glycosylation sites, they must be regarded as potential sites. For this solution, multiple protease digestions coupled with the above LC (CE)–MS–MS strategy will work. (3) The lack of a perfect procedure to collect all glycoproteins. This is an essential issue of glycome projects. In the proposed strategy, we insisted on the use of lectin columns to obtain various sets of glycoproteins. However, this does not necessarily mean that such a combination of lectins, even if their binding specificities are different and

may cover a wide range of glycans, can collect a complete set of glycoproteins. In other words, we may miss some glycopeptides, which pass through any lectin column. From a practical viewpoint, however, there is no material that can adsorb all kind of glycoproteins with no particular specificity. In this context, development of such almighty glyco-adsorbers may be necessary. However, we wonder if it would be biologically meaningful to study glycans that are not recognized by any lectin. Rather, we find a special meaning to use, in particular, endogenous lectins to capture glycoproteins, because such lectins are actually working as decipherers of the glycocode in vivo [4,26].

5. Perspective

Glycomics is an emerging field of post-genome science in this new era. Apparently, glycomics is pioneering new niches of life sciences that nobody has ever approached. Only with the concept of genome–proteome–glycome, can complex life systems be understood, because they consist of a number of cells presenting various messages in the forms of fuzzy, abundant and heterogeneous glycoconjugates. For this achievement, however, development of essential methodologies is absolutely necessary. At the moment, discussion about core strategies for glycomics is still immature, even though we can easily find a number of ($>$150) web sites related to glycome/glycomics. Though still preliminary, we have presented a concept for ''how to approach to glycomics'', as well as a core strategy to realize it. If the two central strategies presented here, i.e. the glyco-catch method and frontal affinity chromatography, are further improved with satisfactory throughput, e.g. a glyco-catch/LC(CE)–MS–MS strategy to identify genes, glycosylation sites, and types of glycans simultaneously, and a glyco-decoder on the basis of a multiple pattern recognition principle using selected members of intelligent lectins, then the individual level of glycomics, like SNPs in genomics, will be realized in the future.

Concurrent glycomics can be categorized into three parts: ''structural glycomics'' occupies the fundamental part, from which diverge ''comparative glycomics'' paying much attention to comparative biochemistry and/or evolutionary science, and ''functional glycomics'' paying much attention to the effects (functions) of glycosylation to elucidate biological meaning of diverse glycans. Presently, these glycomic approaches are almost ''static'' ones, but in future they must evolve into more ''dynamic'' ones, such as those investigating changes in glycosylation states based on a time/video scale observation with relevance to biological significance. Glycomic research has just begun with rather few methods. Success or not greatly depends on development of more efficient, powerful and intelligent technologies. Unless the glycome of many organisms is revealed, the remaining undefined code of life systems, i.e. the glycocode will not be deciphered.

6. Nomenclature

CE	capillary electrophoresis
ConA	concanavalin A
CRD	carbohydrate-recognition domains
EDTA	ethylenediaminetetraacetic acid
FAC	frontal affinity chromatography
GaL6	galectin LEC-6 from the nematode *Caenorhabditis elegans*
GPI	glycosylphosphatidylinositol
HPLC	high-performance liquid chromatography
LC	liquid chromatography
MALDI–TOF	matrix-assisted laser desorption ionization–time of flight
MEPBS	4 m*M* β-mercaptoethanol, 2 m*M* EDTA, 20 m*M* Na phosphate (pH 7.2), 150 m*M* NaCl
MS	mass spectrometry
NHS	*N*-hydroxysuccinimide
PA	pyridylaminated
PBS	phosphate-buffered saline (20 m*M* Na-phosphate, pH 7.2, 150 m*M* NaCl)

Acknowledgements

We thank Drs Kiyoshi Furukawa, Takeshi Sato, Shunji Natsuka and Sumio Hase for their helpful discussions. This work was supported in part by Grants-in-Aid for Scientific Research 13202058 (Priority Area ''Genome Science''), 12680617 and

11771453 from the Ministry of Education, Science, Sports, and Culture of Japan, and by the Mizutani Foundation for Glycoscience.

References

[1] J. Hirabayashi, K. Kasai, Glycoconjug. J. 18 (1999) S33.
[2] J. Hirabayashi, K. Kasai, Trends Glycosci. Glycotechnol. 12 (2000) 1.
[3] J. Hirabayashi, Y. Arata, K. Kasai, Proteomics 1 (2001) 295.
[4] J. Hirabayashi, Y. Arata, K. Kasai, Trends Glycosci. Glycotechnol. 13 (2001) 533.
[5] T. Feizi, Glycoconjug. J. 17 (2000) 553.
[6] N. Taniguchi, A. Ekuni, J.H. Ko, E. Miyoshi, Y. Ikeda, Y. Ihara, A. Nishikawa, K. Honke, M. Takahashi, Proteomics 1 (2001) 239.
[7] H. Kobata, Glycoconjug. J. 17 (2001) 443.
[8] V.C. Washinger, S.J. Cordwell, A. Cerpa-Poljak, J.X. Yan, A.A. Gooley, M.R. Wilkins, M.W. Duncan, R. Harris, K.L. Williams, I. Humphery-Smith, Electrophoresis 16 (1995) 1090.
[9] P.H. O'Farrell, J. Biol. Chem. 250 (1975) 4007.
[10] A. Varki, R. Cummings, H. Freeze, G. Hart, J. Marth (Eds.), Essentials of Glycobiology, Cold Spring Harbor Laboratory Press, New York, 1999.
[11] P. Stanley, E. Ioffe, FASEB J. 9 (1995) 1436.
[12] R.A. Laine, Glycobiology 4 (1995) 759.
[13] J. Hirabayashi, Q. Rev. Biol. 71 (1996) 365.
[14] M. Demetriou, M. Granovsky, S. Quaggin, J.W. Dennis, Nature 409 (2001) 733.
[15] J.S. Griffitts, J.L. Whitacre, D.E. Stevens, R.V. Aroian, Science 293 (2001) 860.
[16] C. Cebo, T. Dambrouck, E. Maes, C. Laden, G. Strecker, J.C. Michalski, J.P. Zanetta, J. Biol. Chem. 276 (2001) 5685.
[17] The *C. elegans* Sequencing Consortium, Science 282 (1998) 2012.
[18] M.D. Adams et al., Science 287 (2000) 2185.
[19] The Arabidopsis Genome Initiative, Nature 408 (2001) 791.
[20] J. Kawai et al., Nature 409 (2001) 685.
[21] International Human Genome Sequencing Consortium, Nature 409 (2001) 860.
[22] N. Tomiya, J. Awaya, M. Kurono, Y. Arata, N. Takahahi, Anal. Biochem. 171 (1988) 73.
[23] N. Takahashi, J. Chromatogr. A 720 (1996) 217.
[24] J. Hirabayashi, Y. Arata, K. Kasai, J. Chromatogr. A 890 (2000) 261.
[25] Y. Arata, J. Hrabayashi, K. Kasai, J. Chromatogr. A 905 (2001) 337.
[26] K. Kasai, J. Hirabayashi, J. Biochem. (Tokyo) 119 (1996) 1.
[27] S. Hase, T. Ikenaka, Y. Matsushima, J. Biochem. (Tokyo) 90 (1981) 407.
[28] Recent Topics on Galectins, J. Hirabayashi (Ed.), Trends Glycosci. Glycotechnol. 9 (45) (1997) 1.
[29] J. Hirabayashi, M. Satoh, K. Kasai, J. Biol. Chem. 267 (1992) 15485.
[30] J. Hirabayashi, T. Ubukata, K. Kasai, J. Biol. Chem. 271 (1996) 2497.
[31] T. Mega, H. Oku, S. Hase, J. Biochem. (Tokyo) 111 (1992) 396.
[32] Y. Ohyama, K. Kasai, H. Nomoto, Y. Inoue, J. Biol. Chem. 260 (1985) 6882.
[33] K.J. Neurohr, N.M. Young, H.H. Mantsch, J. Biol. Chem. 255 (1980) 9205.
[34] N. Kochibe, K. Furukawa, Biochemistry 19 (1980) 2841.
[35] Y. Arata, J. Hirabayashi, K. Kasai, J. Biol. Chem. 276 (2001) 3068.
[36] H. Leffler, S.H. Barondes, J. Biol. Chem. 261 (1986) 10119; Y. Arata, J. Hirabayashi, K. Kasai, J. Biol. Chem. 272 (1997) 26669.
[37] J.H. Neismith, R.A. Field, J. Biol. Chem. 271 (1996) 972.
[38] J. Hirabayashi, K. Kasai, Glycobiology 3 (1993) 297.
[39] D.-I. Liao, G. Kapadia, H. Ahmed, G.R. Vasta, O. Herzberg, Proc. Natl. Acad. Sci. USA 91 (1994) 1428.
[40] Y. Bourne, B. Bolgiano, D.I. Liao, G. Strecker, P. Cantau, O. Herzberg, T. Feizi, C. Cambillau, Nat. Struct. Biol. 1 (1994) 863.
[41] Y.D. Lobsanov, M.A. Gitt, H. Leffler, S.H. Barondes, J.M. Rini, J. Mol. Biol. 233 (1993) 553.
[42] J. Seetharaman, A. Kanigsberg, R. Slaaby, H. Leffler, S.H. Barondes, J.M. Rini, J. Biol. Chem. 273 (1998) 13047.
[43] H. Schachter, Glycoconjug. J. 17 (2000) 465.
[44] F. Sakiyama, T. Masaki, Methods Enzymol. 244 (1994) 126.
[45] H. Nomoto, Y. Inoue, Eur. J. Biochem. 135 (1983) 243.
[46] M.G. Yet, C.C. Chin, F. Wold, J. Biol. Chem. 5 (1988) 111.
[47] S. Suzuki, R. Tanaka, K. Takada, N. Inoue, Y. Yashima, A. Honda, S. Honda, J. Chromatogr. A 910 (2001) 319.
[48] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, J. Mol. Biol. 215 (1990) 403.
[49] N.W. Toribara, J.R. Gum, P.J. Culhane, R.E. Legace, J.W. Hicks, G.M. Petersen, Y.S. Kim, J. Clin. Invest. 88 (1991) 1005.
[50] J. Gonzalez, T. Takao, H. Hori, V. Besada, R. Rodriguez, G. Padron, Y. Shimonishi, Anal. Biochem. 205 (1992) 151.
[51] K. Kasai, Y. Oda, M. Nishikawa, S. Ishii, J. Chromatogr. 376 (1986) 33.
[52] K. Kasai, S. Ishii, J. Biochem. (Tokyo) 84 (1978) 1051.
[53] K. Kasai, S. Ishii, J. Biochem. (Tokyo) 84 (1978) 1061.
[54] Y. Oda, K. Kasai, S. Ishii, J. Biochem. (Tokyo) 89 (1981) 285.
[55] Y. Arata, J. Hirabayashi, K. Kasai, J. Biochem. (Tokyo) 121 (1997) 1002.
[56] M. Malmqvist, Biochem. Soc. Trans. 27 (1999) 335.
[57] C.P. Woudbury Jr., D.L. Venton, J. Chromatogr. B 725 (1999) 113.
[58] D.C. Schriemer, D.R. Bundle, L. Li, O. Hindsgaul, Angew. Chem. Int. Ed. 37 (1998) 3383.
[59] J. Peng, S.P. Gygi, J. Mass Spectrom. 36 (2001) 1083.
[60] S. Suzuki, R. Tanaka, K. Takada, N. Inoue, Y. Yashima, A. Honda, S. Honda, J. Chromatogr. A 910 (2001) 319.
[61] P.K. Jensen, L. Pasa-Tolic, K.K. Peden, S. Martinovic, M.S. Lipton, G.A. Anderson, N. Tolic, K.K. Wong, R.D. Smith, Electrophoresis 21 (2000) 1372.