Glycobiology vol. 20 no. 3 pp. 369–380, 2010
doi:10.1093/glycob/cwp187
Advance Access publication on November 29, 2009

A motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins

Andrew Porter[3,4], Tingting Yue[3,4,6], Lee Heeringa[2,4], Steven Day[5], Edward Suh[5], and Brian B Haab[1,4]

[4]Van Andel Research Institute, Grand Rapids, MI 49503; [5]Translational Genomics Institute (TGen), Phoenix, AZ 85004; and [6]Cell and Molecular Biology Program, Michigan State University, East Lansing, MI 48823, USA

Received on October 2, 2009; revised on November 17, 2009; accepted on November 22, 2009

Glycan arrays have enabled detailed studies of the specificities of glycan-binding proteins. A challenge in the interpretation of glycan array data is to determine the specific features of glycan structures that are critical for binding. To address this challenge, we have developed a systematic method to interpret glycan array data using a motif-based analysis. Each glycan on a glycan array is classified according to its component sub-structures, or motifs. We analyze the binding of a given lectin to each glycan in terms of the motifs in order to identify the motifs that are selectively present in the glycans that are bound by the lectin. We compared two different methods to calculate the identification, termed intensity segregation and motif segregation, for the analysis of three well-characterized lectins with highly divergent behaviors. Both methods accurately identified the primary specificities as well as the weaker, secondary specificities of all three lectins. The complex binding behavior of wheat germ agglutinin was reduced to its simplified, independent specificities. We compiled the motif specificities of a wide variety of plant lectins, human lectins, and glycan-binding antibodies to uncover the relationships among the glycan-binding proteins and to provide a means to search for lectins with particular binding specificities. This approach should be valuable for rapidly analyzing and using glycan array data, for better describing and understanding glycan-binding specificities, and as a means to systematize and compare data from glycan arrays.

Keywords: glycan arrays/glycan microarrays/glycan-binding protein/lectin/motif analysis

Introduction

Glycan-binding proteins are important both for their biological functions and for their use as analytical reagents. Proteins that specifically recognize and interact with carbohydrates, called lectins, are found in every type of known organism and play major roles in biological processes such as immune recognition and regulation, inflammatory responses, cytokine signaling, and cell adhesion (Varki et al. 1999). Lectin interactions with their carbohydrate ligands also contribute to various pathologies (Dennis et al. 1999; Dube and Bertozzi 2005; Fuster and Esko 2005; Lau and Dennis 2008) and form the basis of multiple congenital disorders (Freeze 2006; Freeze and Aebi 2005). As analytical reagents, lectins and glycan-binding antibodies are extremely valuable for detecting and isolating specific glycans (Hirabayashi 2004; Sharon 2007). They have been used in diverse experimental formats, such as immunohistochemistry (Satomura et al. 1991; Osako et al. 1993), affinity electrophoresis (Shimizu et al. 1996), immunofluorescence cell staining (Wearne et al. 2006), lectin arrays (Angeloni et al. 2005; Kuno et al. 2005; Pilobello et al. 2005), and antibody (Chen et al. 2007; Yue et al. 2009) and protein arrays (Patwa et al. 2006; Li et al. 2009), to characterize both normal and pathological glycosylation. A critical step in understanding the biology of lectins and in using them as analytical reagents is to characterize their glycan-binding specificities. The glycan-binding specificities of many lectins have been well characterized, but many others remain for which little is known. Improved methods of systematically analyzing and categorizing glycan binding specificities are needed.

The specificities of glycan-binding proteins are typically determined through measuring the binding levels to a wide variety of isolated glycan structures, using methods such as frontal affinity chromatography (Hirabayashi et al. 2002, 2003; Hirabayashi 2004; Tateno et al. 2007) and glycan microarrays (Wang 2003; Culf et al. 2006). An advantage of glycan microarrays over chromatography methods is the use of minimal amounts of glycans to probe numerous interactions, which is significant considering the time and expense involved in synthesizing glycans. Several related glycan microarray technologies have been developed, with diversity in the surface and attachment chemistries, the types of glycans used on the arrays, and the methods of detecting binding to the glycans on the microarrays (Wang 2003; Culf et al. 2006). The availability of glycan microarray technology and its associated data has been greatly increased through the Consortium for Functional Glycomics (CFG), which provides microarrays containing over 300 biologically relevant, synthesized glycans (Blixt et al. 2004) to participating researchers and makes the data publicly available. The data from multiple plant lectins, animal lectins, and glycan-binding antibodies have been assembled and made available on the CFG website. This expanding availability of glycan microarray data presents an opportunity for increasing the knowledge of the specificities of glycan-binding proteins.

A current limitation in making full use of glycan microarray data is the lack of systematic analysis methods for extracting information. Systematic analysis methods are necessary because

[1]To whom correspondence should be addressed: Tel: +1-616-234-5268; Fax: +1-616-234-5269; e-mail: Brian.Haab@vai.org
[2]Present address: College of Human Medicine, Michigan State University, East Lansing, MI 48824, USA.
[3]These authors contributed equally to this work.

of the nature of glycan microarray data. Because of the structural complexity of some oligosaccharides, and because certain lectins may have multiple, related specificities, the task of sifting through glycan array data to discern binding specificities can be difficult and time-consuming. An analytical tool for determining binding specificities from glycan array data could ease this task as well as add definable and quantifiable interpretation to the data. In addition, the ability to automate glycan array analysis would enable the cataloging and comparisons of many datasets, which could be used for searching and higher-level analyses.

A strategy for discerning lectin specificities is to define the component parts, or motifs, of oligosaccharides that are responsible for lectin binding. Such a strategy was used in a frontal affinity chromatography study to characterize the specificities of two different plant lectins for core α1,6-linked fucose (Tateno et al. 2009). A visual inspection of the binding levels to glycans containing specific features revealed this specificity. We have expanded on that approach by defining numeric statistics to describe the preference of a lectin for particular motifs. In this paper, we describe two different analytical approaches for determining binding specificities of lectins and characterize their performance on three lectins with greatly divergent behaviors. We further show the value of the method for systematizing and comparing the specificities of multiple plant lectins, animal lectins, and glycan-binding antibodies. The information from these analyses can be used through a searchable database that is now available.

Results

Glycan array data

We obtained publically available glycan array data from the Consortium for Functional Glycomics (CFG). The experimental steps in the generation of glycan array data were described in detail previously (Blixt et al. 2004) and are briefly summarized in the methods section. Data from four versions of the CFG array were obtained. Each successive array version contained an increasing number of glycans, from 266 glycans in version 2.0 to 377 glycans in version 3.1. The experiments and primary analyses were performed by the CFG, using glycan-binding proteins (lectins and glycan-binding antibodies) that were supplied by participating investigators. Each glycan-binding protein was incubated on an array, and the relative binding levels to the various glycans on the arrays were measured by fluorescence scanning.

The goal of this work was to develop an automated method to extract from the glycan array data information about the binding specificity of each glycan-binding protein. For each glycan array experiment, the fluorescence signal intensity measured at each glycan reflects the amount of lectin or antibody binding to that glycan (Figure 1). By visual inspection of the data from certain lectins, common features among the bound glycans are readily discernible. For example, the lectin VVL strongly binds glycans containing terminal GalNAc and shows very little binding to others (Figure 1A). The binding of ConA to terminal mannose-containing structures is clear (Figure 1B), although weaker-affinity binding to other structures also is apparent. For other lectins, the visual interpretation of the information can be difficult, for example if many structures are bound or if broader specificities exist, as with WGA (Figure 1C). The task of discerning common features among multiple, complex glycan structures becomes very difficult once the specificity involves non-terminal structures or multiple glycan structures. A systematic method of analyzing glycan array data would provide a more objective approach to determining binding specificities and would enable more rapid and rigorous comparisons of multiple datasets.

Motif-based analyses of binding specificities

To address this problem, we began by converting the glycan-binding information from the arrays to motif-based information. We created a list of 63 motifs that regularly appear in biology and that represent functionally important sub-structures found in larger carbohydrate chains—motifs such as lactosamine, terminal α1,4-linked fucose, and terminal mannose. Next, we recorded the presence or absence of each motif on each of the complex glycan structures on the glycan array (Figure 2A). A description of the definitions of the motifs is found in Table I, and the complete tables of motifs and glycan structures for each of the four array versions are provided in supplementary Tables I–IV. Some motifs were present on many glycans (lactosamine was on 93 glycans in array version 2.0), and others on just a few (Lewis B was on just one glycan in version 2.0). supplementary Figure 1 shows the broad range of representation of the motifs on each array version. In terms of the number of motifs found on each glycan, every glycan was represented by at least one motif, except for a single glycan found on array versions 2.1 and higher-alpha-linked rhamnose—for which we did not define a motif. The glycans containing only one motif were generally mono- or di-saccharides. Most glycans contained more than one motif, up to a maximum of 12 motifs found on two complex glycans. Therefore, the 63 motifs chosen here provided a useful "vocabulary" to describe the glycans on these arrays.

The conversion from glycans to motifs allowed us to determine which motifs could be important for the binding specificity of a given lectin, based on which motifs were specifically enriched among glycans that were bound by that lectin. Two complementary strategies for that determination were tested. The first approach, which we called the intensity segregation method (Figure 2B), was to segregate the glycans by fluorescence intensity, corresponding to glycans that were bound or not bound by the lectin, followed by determining which motifs were largely present in the high-intensity group but not the low-intensity group. For each motif, we subtracted the percentage of glycans in the unbound group containing the motif from the percentage of glycans in the bound group containing the motif. A high positive value indicates a high enrichment of the motif in the bound group of glycans, whereas a high negative value indicates the opposite. The second approach, which we called the motif segregation method (Figure 2C), was to segregate the glycans according to the presence or absence of a given motif, and then to calculate the statistical difference in fluorescence signal between the glycans in the two groups (Figure 2C). We used the Mann–Whitney test for the comparison because the intensities might not be normally distributed or precisely quantitative. For graphing and comparison purposes, we log-transformed the *P*-value and multiplied it by the sign of the *z*-score, so that a high positive value indicates a strong association with high lectin binding and a high negative value indicates the opposite

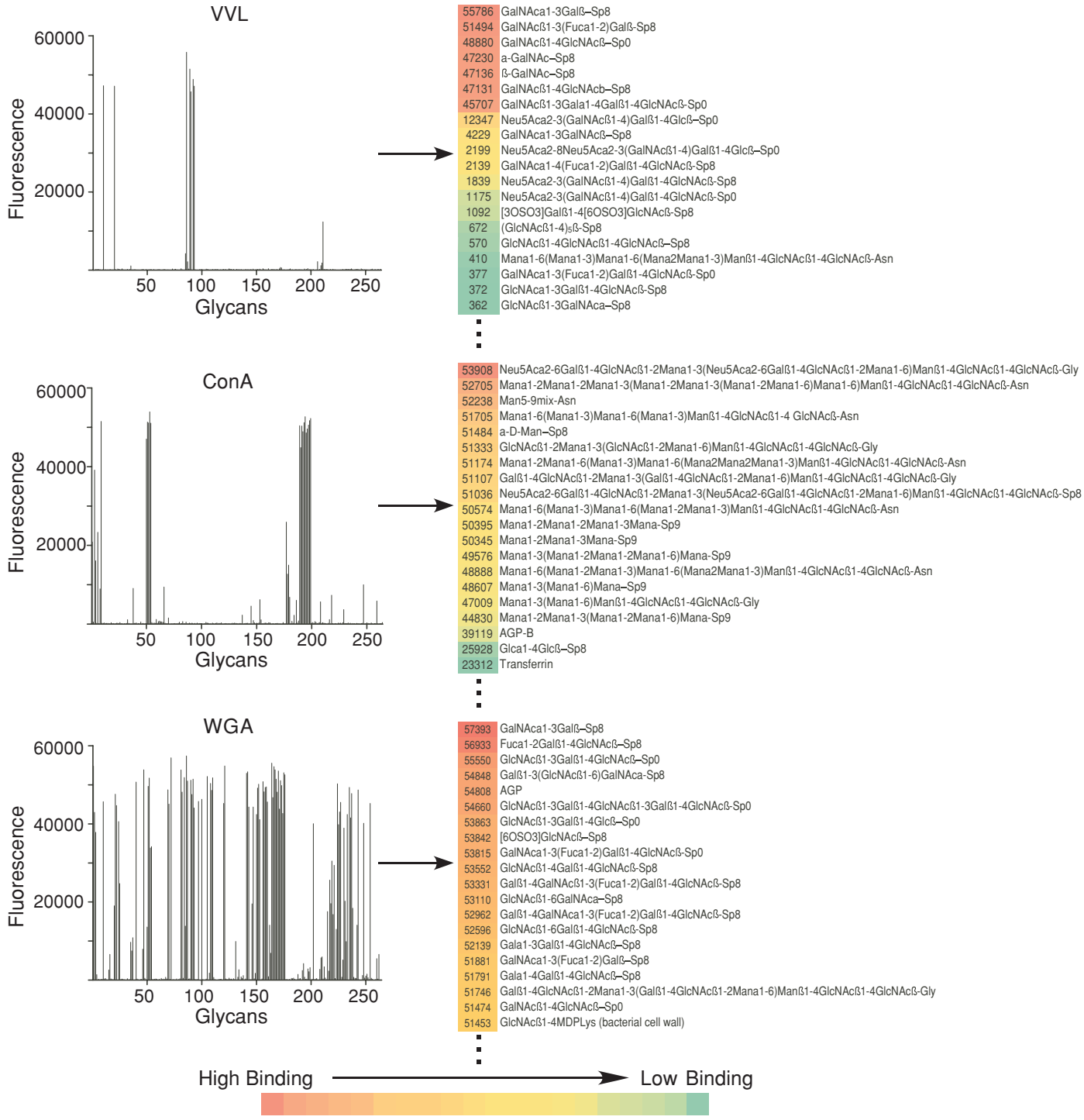

Fig. 1. Glycan array data. The plot represents the fluorescence intensities of individual glycans (ordered along the *x*-axis) on glycan arrays, after incubation with either VVL (top), ConA (middle), or WGA (bottom). Next to each graph, the top-ranking glycans are listed along with the numeric value of the corresponding fluorescence intensities.

association. These two approaches represent the two broad categories of making associations between binding levels and motifs; other statistical tests could be applied within this framework once the glycans or motifs are segregated.

We compared the performance of these two strategies for the three lectins highlighted above (Figure 3). These three lectins represent distinct types of binding patterns on the arrays, and so provide a broad test and basis for comparing the two approaches. For VVL, both methods ranked terminal beta-linked GalNAc far above the rest, which agrees with the known specificity of VVL, but only the intensity segregation method scored terminal alpha-linked GalNAc high. For ConA, both methods

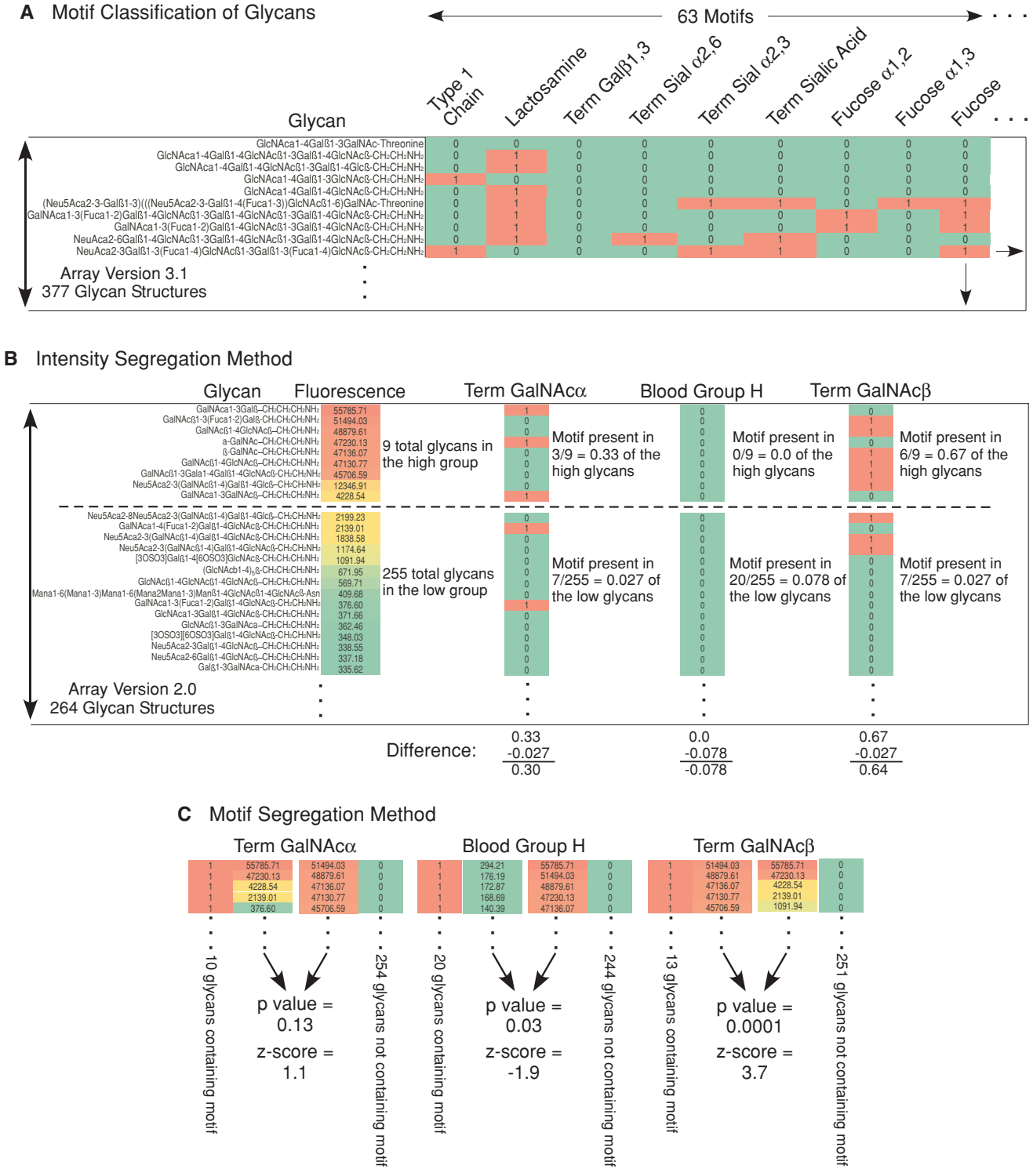

Fig. 2. Motif-based analysis of glycan-binding specificities. (**A**) Classifying glycans by their component motifs. For each glycan, the presence or absence of each of 63 possible motifs was recorded with a "1" or "0", respectively. (**B**) The intensity segregation method for identifying motif specificities. For each set of glycan array data, a threshold is set which segregates the glycans with high intensity from those with low intensity. The thresholds were based on the distributions of the fluorescence intensities in order to maximize segregation between low-intensity and high-intensity spots. For each motif, the percent presence is calculated in both the high-intensity glycans and the low-intensity glycans, and the difference between the two fractions is calculated. The example here shows the analysis of glycan-array data from the lectin VVL for three different motifs. (**C**) The motif segregation method. For each motif, the glycans are segregated according to the presence or absence of that motif. A statistical test compares the intensities or ranks of the glycans containing the motif to the glycans not containing the motif. The example here shows data from VVL, and *P*-values and *z*-scores based on the Mann–Whitney test.

Table I. The motifs and their definitions.

Motif number	Motif	Definition
1	Terminal Galβ1,3	Terminal Galactose with a β1,3 linkage to the proximal glycan
2	Terminal Galβ1,4	Terminal Galactose with a β1,4 linkage to the proximal glycan
3	Terminal Fucα1,2	Terminal fucose with an α1,2 linkage to the proximal glycan; can include coming off of a branch that extends further
4	Terminal Fucα1,3	Terminal fucose with an α1,3 linkage to the proximal glycan; can include coming off of a branch that extends further
5	Terminal Fucα1,4	Terminal fucose with an α1,4 linkage to the proximal glycan; can include coming off of a branch that extends further
6	Terminal Fucα1,6	Terminal fucose with an α1,6 linkage to the proximal glycan; can include coming off of a branch that extends further
7	Terminal Fuc	Terminal fucose of any type, including α and β, including a fucose that comes off of a branch that extends further
8	Terminal Sial α2,3	Terminal sialic acid with an α2,3 linkage to the proximal glycan; can include Neu5Ac, Neu5Gc, and KDN
9	Terminal Sial α2,6	Terminal sialic acid with an α2,6 linkage to the proximal glycan; can include Neu5Ac, Neu5Gc, and KDN
10	Terminal Sial β2,6	Terminal sialic acid with an β2,6 linkage to the proximal glycan; can include Neu5Ac, Neu5Gc, and KDN
11	Terminal Sial α2,8	Terminal sialic acid with an α2,8 linkage to the proximal glycan; can include Neu5Ac, Neu5Gc, and KDN
12	Terminal Sialic Acid	Terminal sialic acid with an α or β linkage to the proximal glycan; can include Neu5Ac, Neu5Gc, and KDN
13	Neu5Gc	Terminal Neu5Gc with an α or β linkage to the proximal glycan
14	Sialylated Tn	A sialic acid (Neu5Ac, Neu5Gc, or KDN) bound to a GalNAc in any way, no distinction between α, β, or the numeration of the bond; multiple sialic acids may be bound to the GalNAc, but Gal may not be bound to GalNAc
15	Sialylated T-antigen	A sialic acid (Neu5Ac, Neu5Gc, or KDN) bound to a GalNAc with either an α or β linkage, and a galactose β1,3 linked to the GalNAc; the Gal branch may have glycans following it
16	9NAcNeu5Ac	Terminal 9NAcNeu5Ac linked to any other unit
17	Terminal Lactosamine	A terminal lactosamine unit, consisting of a terminal Galβ1,4GlcNAcβ, with no distinction of the numeration of the β-linkage from the GlcNAc to the proximal glycan
18	Internal Lactosamine	An interminal lactosamine unit, consisting of a Galβ1,4GlcNAcβ, with no distinction of the numeration of the β-linkage from the GlcNAc to the proximal glycan or what follows the galactose
19	Branching	A GlcNAcβ1,6 branch from GalNAc or GlcNAc; all branches can have more glycans following the initial glycans
20	Type1 Chain	A neolactosamine unit, present anywhere in the chain, consisting of terminal or internal Galβ1,3GlcNAcβ, with no distinction of the numeration of the β-linkage from the GlcNAc to the proximal glycan
21	Lactosamine	A lactosamine unit, present anywhere in the chain, consisting of terminal or internal Galβ1,4GlcNAcβ, with no distinction of the numeration of the β-linkage from the GlcNAc to the proximal glycan
22	I-blood group antigen (GlcNAcβ1,6Gal)	A GlcNAcβ1,6 linked to a galactose anywhere in the chain; other glycans may be attached to the galactose also and the chain may continue past GlcNAc
23	Polylactosamine	A chain of lactosamine units, consisting of a two or more consecutive Galβ1,4GlcNAcβ1,3 units
24	Neo-polylactosamine	A chain of neolactosamine units, consisting of a two or more consecutive Galβ1,3GlcNAcβ1,3 units
25	Lewis X	A unit of Galβ1,4(Fucα1,3)GlcNAc present in the chain, either in a terminal or internal position, except in the cases of being the base structure for Lewis y, sialyl Lewis x, 3′ sulfo Lewis x, or 6 sulfo-sialyl Lewis x
26	Lewis Y	A unit of Fucα1,2Galβ1,4(Fucα1,3)GlcNAc present in the chain, either in a terminal or internal position
27	Sialyl Lewis X	A unit of Sialic Acidα2,3Galβ1,4(Fucα1,3)GlcNAc present in the chain, either in a terminal or internal position, with the sialic acid being either Neu5Ac, Neu5Gc, or KDN
28	3′Sulfo Lewis X	A unit of [3OSO3]Galβ1,4(Fucα1,3)GlcNAcβ present in the chain, either in a terminal or internal position, note the 3′sulfate is on galactose, not GlcNAc
29	6′Sulfo-sialyl Lewis X	A unit of sialic acidα2,3Galβ1,4(Fucα1,3)(6OSO3)GlcNAcβ present in the chain, either in a terminal or internal position, with the sialic acid being either Neu5Ac, Neu5Gc, or KDN; note the 6′sulfate is on galactose, not GlcNAc
30	Lewis A	A unit of Galβ1,3(Fucα1,4)GlcNAc present in the chain, either in a terminal or internal position, except in the cases of being the base structure for Lewis b, sialyl Lewis a, or 3′sulfo Lewis a
31	Lewis B	A unit of Fucα1,2Galβ1,3(Fucα1,4)GlcNAc present in the chain, either in a terminal or internal position
32	Sialyl Lewis A	A unit of Sialic Acidα2,3Galβ1,3(Fucα1,4)GlcNAc present in the chain, either in a terminal or internal position, with the sialic acid being either Neu5Ac, Neu5Gc, or KDN
33	3′Sulfo Lewis A	A unit of [3OSO3]Galβ1,3(Fucα1,4)GlcNAcβ present in the chain, either in a terminal or internal position, note the 3′sulfate is on galactose, not GlcNAc
34	SO4	A sulfate group is present at least once on any glycan, in any position
35	*O*-Glycan Core 1	A Galβ1,3GalNAc unit is present either as a base structure or in a terminal position; noted still when a sialylated T-antigen is present and when the base has other glycan additions except in the cases of being core 2, 3, or 4
36	*O*-Glycan Core 2	A Galβ1,3(GlcNAcβ1,6)GalNAc unit is present either as a base structure or in a terminal position; includes the structure when other glycans are added to this base
37	*O*-Glycan Core 3	A GlcNAcβ1,3GalNAc unit is present either as a base structure or in a terminal position; noted still when the base has other glycan additions except in the case of being core 4
38	*O*-Glycan Core 4	A GlcNAcβ1,3(GlcNAcβ1,6)GalNAc unit is present either as a base structure or in a terminal position; includes the structure when other glycans are added to this base
39	*N*-Glycan high-mannose	A glycan chain with a Manα1-3(Manα1-6(Manα1-3)Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ base; other mannose glycans may be added to this base, but no other glycans can be added
40	*N*-Glycan hybrid	A glycan chain with a Mana1-3(Mana1-6)Manb1-4GlcNAcb1-4GlcNAcb base with GlcNAcb1-2 linked to only one of the two terminal mannose glycans (Mana1,3 or Mana1,6); the branch with the GlcNAcb1,2 can continue to grow with any glycan, but only mannose may be present on the other mannose
41	*N*-Glycan complex	A glycan chain with a GlcNAcb1-2Mana1-3(GlcNAcb1-2Mana1-6)Manb1-4GlcNAcb1-4GlcNAcb base; further additions of any glycans may be made at the mannose (Manb1,4) and beyond, fucose may be added to the first GlcNAc
42	Truncated *N*-glycan/precursors	A glycan chain that at least contains GlcNAcb1-4GlcNAcb at the base; frequently includes two mannose units
43	Terminal Galα	Terminal galactose alpha-linked to any other group

(*Continued*)

Table I. (Continued)

Motif number	Motif	Definition
44	Terminal Galβ	Terminal galactose beta-linked to any other group
45	Terminal GalNAcα	Terminal *N*-acetylgalactosamine alpha-linked to any other group
46	Terminal GalNAcβ	Terminal *N*-acetylgalactosamine beta-linked to any other group
47	Terminal Manα	Terminal mannose alpha-linked to any other group
48	Terminal Manβ	Terminal mannose beta-linked to any other group
49	Terminal Glcα	Terminal glucose alpha-linked to any other group
50	Terminal Glcβ	Terminal glucose beta-linked to any other group
51	Terminal GlcNAcα	Terminal *N*-acetylglucosamine alpha-linked to any other group
52	Terminal GlcNAcβ	Terminal *N*-acetylglucosamine beta-linked to any other group
53	Terminal GlcA Glucuronic acid	Terminal glucoronic acid linked to any other group
54	Blood A antigen	A glycan chain with a GalNAca1-3(Fuca1-2)Galb1-3 unit, including both as a terminal unit and internal with further extensions
55	Blood B antigen	A glycan chain with a Gala1-3(Fuca1-2)Galb1-3 unit, including both as a terminal unit and internal with further extensions
56	Blood H antigen	A glycan chain with a Fuca1-2Galb1-3 unit, including both as a terminal unit and internal with further extensions
57	Pk antigen	A glycan chain with Gala1-4Galb1-4Glcb without the addition of any glycan extensions
58	P antigen	A glycan chain with GalNAcb1-3Gala1-4Galb1-4Glcb as a base
59	P1 antigen	A glycan chain with Gala1-4Galb1-4GlcNAcb1-3Galb1-4Glcb as a base; anything could be extended from it
60	Terminal Neu5Acα2,3Gal	Terminal Neu5Aca2,3Gal linked to any other unit
61	Terminal Neu5Acα2,6Gal	Terminal Neu5Aca2,6Gal linked to any other unit
62	Terminal Neu5Acα2,3GalNAc	Terminal Neu5Aca2,3GalNAc linked to any other unit
63	Terminal Neu5Acα2,6GalNAc	Terminal Neu5Aca2,6GalNAc linked to any other unit

correctly predicted high specificity for terminal alpha-linked mannose and weaker specificity for terminal glucose, and the motif segregation method predicted very weak significance for α1,3-linked fucose. For WGA, both methods gave high scores for its known specificities of internal lactosamine, terminal GlcNAc, and Neu5Acα2,3Gal, with some differences between the methods in the weaker specificities.

These results show that both methods accurately score the primary specificities of these lectins. This accuracy likely extends to other lectins, since these three lectins span a variety of behaviors, including strong specificity for a small number of elements, as with VVL, or broader specificities, such as ConA and especially WGA. However, the two methods showed differences in their scoring of the weaker specificities. The differences may be due to susceptibilities to inaccuracies in the weaker associations inherent to each method. For example, the rank-based Mann–Whitney test used in the motif segregation method could pick up weakly significant associations that might not be truly meaningful based on signal intensities, such as the weakly significant sialyl Lewis X motif found for VVL (Figure 3A) and the fucose α1,3 motif found for ConA (Figure 3B). The use of another statistical comparison such as the *t*-test with the motif segregation method could provide more information about the nature of differences in the signal intensities between the groups of glycans. Therefore, the intensity segregation method may have an advantage in clearly defining motif enrichment in a high-binding group, without respect to the quantitation of the signal intensities, which can be highly variable in glycan array data. However, the results of the intensity segregation method depend highly on the threshold used to define the high and low signals, which can be difficult to define in certain situations. Since the two methods agree and are accurate for the primary specificities, we chose the motif segregation method for subsequent analyses because of its relatively certainty and ease of automation, while recognizing that the use of two or three different statistical comparisons may be necessary to gain a more complete picture of what associations are truly meaningful.

In finding motifs that are associated with a stronger signal, it is useful to determine whether particular motifs are independently meaningful or rather are highly correlated with other motifs. This question relates to whether a motif truly is the binding determinant or is simply structurally correlated with the binding determinant. An examination of how the top-scoring motifs correlate with each other and with the signal intensities of the glycans on which they are found provides insights into this question (Figure 4). For VVL (Figure 4A), the two top motifs, GalNAcα and GalNAcβ, do not correlate with each other or with any other motif, but they are separately found on all the glycans with high signal intensity. This finding indicates that these two motifs are independent and are the only binding determinants on the array. For ConA (Figure 4B), the mannoseα motif independently is present on a large group of the high-intensity glycans, while the complex *N*-glycan motif is independently found on another group, indicating that ConA has specificity for both. A group of weaker-intensity glycans independently contain alpha-linked glucose, a known weak ligand of ConA, but no other motif shows independent associations.

WGA gives a more complex picture (Figure 4C). A number of motifs are found in the high-intensity glycans, and it is difficult by visual inspection to determine which are independent. It is apparent that terminal GlcNAc and lactosamine are not correlated with each other, but the relationship among the other motifs is not clear. A method to gain insight into the relationships among the motifs and glycans is to cluster both the motifs and glycans by similarity (Figure 4D). This cluster reveals that WGA binds independent groups of glycans that are defined by certain dominant motifs. Lactosamine (Galβ1,4GlcNAc) is a major group, including the internal, terminal, or sialylated versions. Terminal GlcNAc forms another independent group. GlcNAc in general is not a binding determinant, since other motifs containing GlcNAc, such as the Lewis antigens and the type-1 chain

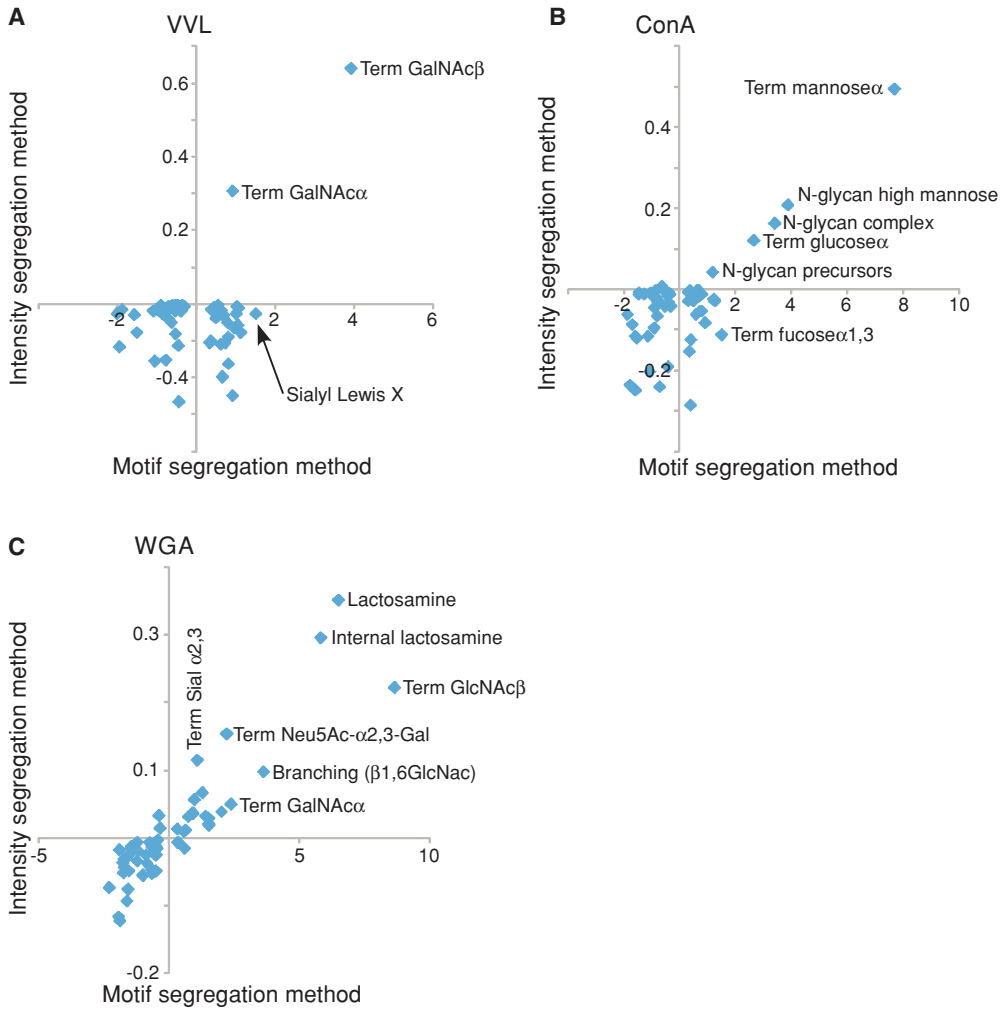

Fig. 3. Comparisons of the intensity segregation and motif segregation methods. The scores for each motif derived from the intensity segregation method are plotted with respect to the scores from the motif segregation method for the lectins (**A**) VVL; (**B**) ConA; and (**C**) WGA. Each data point is a separate motif. The intensity segregation scores are the fractional differences calculated as shown in Figure 2B, and the motif segregation scores are the logged (base 10) Mann–Whitney *P*-values calculated as in Figure 2**C**. The *P*-values were multiplied by the sign of the *z*-score, so that negative values indicate motifs negatively associated with lectin binding.

(Galβ1,3GlcNAc), did not have significant scores. Some motifs are subsets of the lactosamine motifs, such as polylactosamine, branching, and *O*-glycan Core 2, and so are likely themselves not the actual binding determinants. However, other motifs are present in groups independent from terminal GlcNAc and lactosamine, such as alpha2,3-linked sialic acid and alpha- and beta-linked terminal GalNAc, so these motifs represent additional specificities of WGA. Therefore, WGA has the dominant specificities of GlcNAc and lactosamine with some additional specificities. This analysis shows the value of the motif-based method for bringing order to complex glycan-binding behaviors.

Comparative motif specificities of sets of glycan-binding proteins

Once an automated method is in place for scoring the binding specificities of lectins using glycan array data, the data from multiple experiments can be rapidly analyzed and probed as a group. We examined the relationships between the motif specificities of a group of 84 plant lectins (Figure 5). The lectins clustered in distinct groups that in most cases were identifiable by a corresponding group of motifs with high-significance scores. The clearest groups of lectins were defined by the terminal Gal β1,3, terminal Gal β1,4 terminal GlcNAc, sialic acid α2,6, fucose, mannose, and internal lactosamine motifs. Most of the lectins with well-characterized specificities fall within their expected groupings. For example, the lectins PNA and BPL are binders of terminal α1,3-linked Gal and terminal Gal, respectively; AAL binds multiple linkages of terminal fucose; and ConA binds high-mannose and complex *N*-glycans. The areas of highly negative scores that indicate the certain motifs are almost never bound by particular lectins. Most notably, α2,3-linked and α2,6-linked sialic acid each have associated groups of lectins with strong negative scores, probably because the negatively charged capping group can prevent binding to the underlying saccharides. Therefore, in addition to the motif preferences of lectins, the motifs that are not preferred or are inhibitory also can be found using this tool. The combined analysis of multiple motifs and lectins may provide a useful means

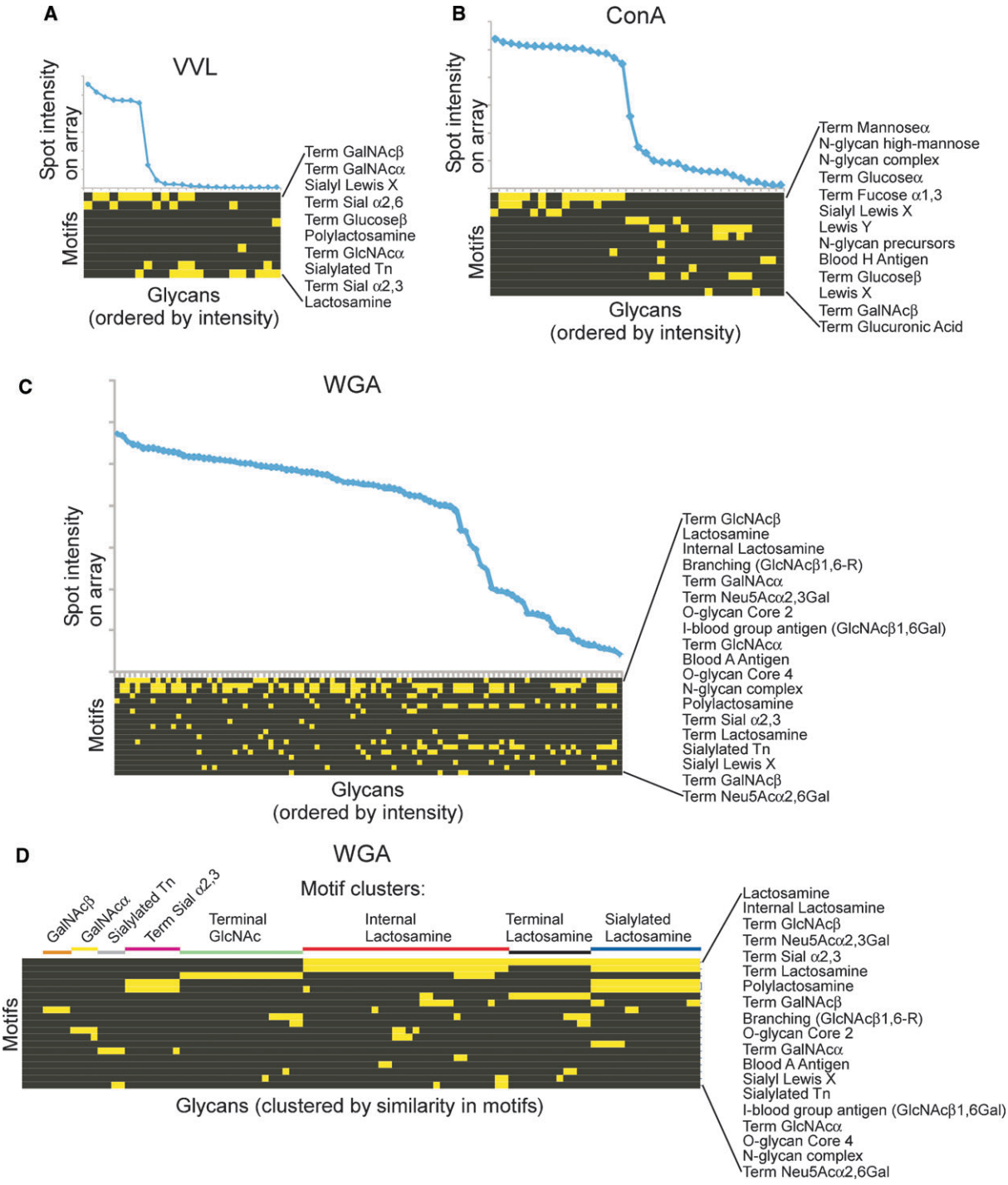

Fig. 4. Associations between motifs that are enriched in high-intensity glycans. For the lectins VVL (**A**), ConA (**B**), and WGA (**C**), the top glycans were ordered by intensity, and the presence or absence of the top-ranking motifs (by motif segregation) is indicated for each glycan by yellow or black squares, respectively. For VVL, GalNAcα was placed as the second-ranking motif for clarity, although it was not ranked second by motif segregation. All other motifs are ordered from top to bottom by their motif segregation *P*-value. (**D**) Groupings among the glycans containing the top-ranking motifs. The data from (**C**) were clustered by similarity among both the glycans (columns) and the motifs (rows). The glycans group according to the dominant and independent presence of distinct motifs, as labeled above the cluster.

of classifying the binding behavior of glycan-binding proteins for which little is known.

We also applied this method to the analysis of animal lectins and glycan-binding antibodies. The animal lectins show many of the same groupings of specificities, such as lactosamine, terminal galactose, but with some differences, such as more lectins that bind α2,3-linked sialic acid and fewer that bind terminal mannose (supplementary Figure 2). The organization of this information is the first step in the probing and further study of these specificities. The motif specificities of the glycan-binding

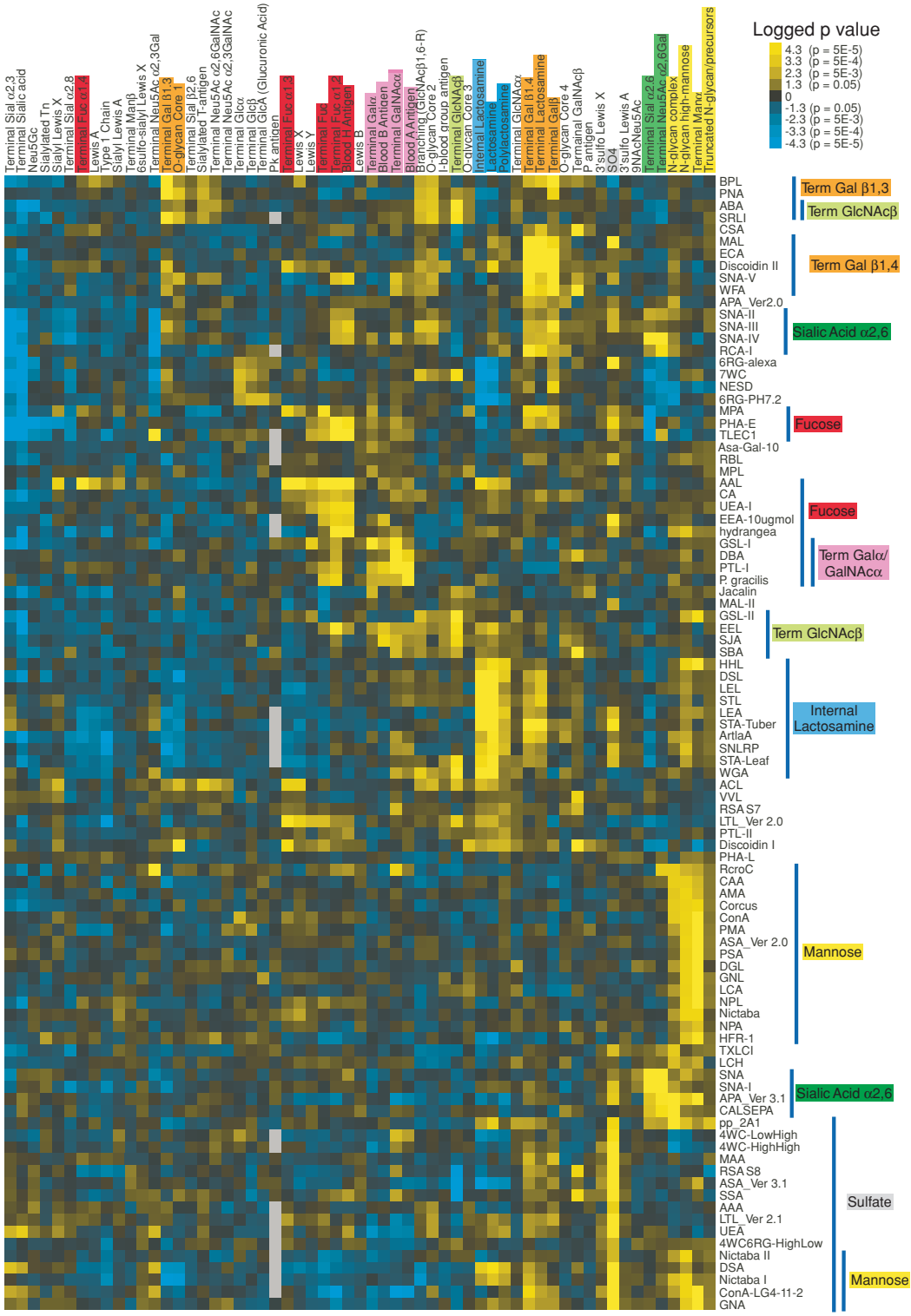

Fig. 5. Comparative analysis of the motif specificities of plant lectins. The data from 113 different glycan array incubations of plant lectins were analyzed using the motif segregation method, and the logged *P*-values (multiplied by the sign of the *z*-score) are clustered by similarity among both the motifs (columns) and the rows (lectins). Eighty-four unique lectins are presented, with some lectins run on multiple arrays under varying conditions. The *P*-values are indicated by the color bar. Clusters of motifs and lectins with high-significance scores are highlighted.

antibodies showed that some antibodies are highly specific for their intended target, while others have either low specificity for any motif or discernable cross reactivity with related motifs (supplementary Figure 3). This conclusion about glycan-binding antibodies also was determined in a previous study of glycan array data (Manimala et al. 2007).

This analysis also provides a tool to search for lectins with particular specificities. Once the data are assembled in a systematic way such as this, one may probe the data to identify a lectin that specifically binds certain motifs without binding others. For example, the lectin ConA is widely used as an analytical reagent because of its high affinity, but its specificity is quite broad, since it binds both high-mannose and complex *N*-glycans as well as glucose (Figures 3 and 5). Using the data represented in Figure 5, one may search for lectins that bind particular types of *N*-glycans, such as high-mannose *N*-glycans, without detecting complex-type *N*-glycans or any other motif. Among the lectins in the mannose-binding cluster of Figure 5, some bind high-mannose *N*-glycans and terminal mannose but do not strongly bind anything else, such as HFR-1 (Hessian fly response 1). We have developed an online searchable database that can be used to mine these analyses, available through the external links page at the CFG website (www.functionalglycomics.org).

Discussion

The characterization of the specificities of glycan-binding proteins is of primary importance in understanding the biological roles of protein–glycan interactions and for using glycan-binding proteins as analytical reagents. The glycan array has been an important tool for such investigations, but the full use of glycan array data has not yet been realized due to the lack of automated and systematized methods for extracting information. We demonstrate here the development of a motif-based analysis of glycan-array data to address that need. We tested two different approaches to calculating the contribution of motifs to binding, intensity segregation and motif segregation, and we showed that both methods can accurately identify the primary specificities of lectins with a variety of binding behaviors. The application of the method to multiple lectins and glycan-binding antibodies allowed an analysis of the relationships among glycan-binding proteins and a means of searching for proteins with specific binding properties. This tool should facilitate the rapid analysis and use of glycan microarray data and will enable comparative analyses and searches of the existing data.

The testing of the intensity segregation and the motif segregation methods on lectins with highly divergent behaviors (VVL, ConA, and WGA) provided useful information about the performance, limitations, and possible improvements of the method. Both methods performed well in picking up the primary, known specificities of each lectin, showing the inherent reliability of the general approach. An area for improvement was revealed by the fact that the motif segregation method did not perform as well as the intensity segregation method in picking up the known specificity of VVL for GalNAcα. The intensity segregation method found the enrichment of the three out of 10 GalNAcα-containing glycan in the nine total glycans above the threshold (Figure 2B). Since the other seven GalNAcα-containing glycans had low ranks among the glycans on the array, the motif segregation method found no overall contribution of this motif to VVL binding. An examination of the GalNAcα-containing glycans revealed that all the low-intensity glycans contained fucose α1,2-linked to the proximal Gal, which is characteristic of the blood group A/B/O antigens. At least in this setting, the proximal fucose α1,2 motif appears to be detrimental to VVL binding. More experimentation would be required to make conclusions, but this analysis shows the potential of the method to pull out contributions from previously unidentified structural features.

The above observation also highlights the fact that some binding determinants may be more complex than the relatively simple motifs defined here, especially if they span larger structures with physically separated contact points. An example of this type of behavior is the lectin *Pisum sativum*, which shows primary specificity for core fucose (fucose that is α1,6-linked to the base on an *N*-glycan) but also is affected by the branching structures of the associated *N*-glycan (Tateno et al. 2009). One approach to handle that complexity is to define additional, more detailed motifs. Alternatively, combinations of the existing motifs might reveal more complex binding specificities. Methods could be used that have been developed for the classification of patient samples using combinations of gene expression profiles from DNA microarray data, such as a class prediction method (Golub et al. 1999). Further analyses will be required to determine which approaches most accurately reveal complex binding determinants.

Other analytic methods may be valuable within this general approach. Modifications to the motif segregation method that are more selective in the glycans that are compared may be useful. For example, instead of comparing all glycans containing a motif to all glycans not containing the motif, it may be better to just compare structurally similar glycans that either contain or do not contain the motif. An example would be to compare glycans containing Neu5Acα2,3Galβ1,4 only to glycan containing terminal Galβ1,4 (instead of comparing to all other glycans) in order to test the Neu5Acα2,3 motif. Such a strategy would minimize the potential skewing of results by greatly unbalanced representations of certain motifs on the arrays and would more directly test the importance of specific motifs. We are currently exploring approaches built on that concept.

The accuracy of the analysis naturally depends on the inherent accuracy of the glycan array data. Glycan arrays may be susceptible to nonspecific binding, and certain binding interactions may be missed due to valency or orientation requirements that are only present when glycans are supported on the appropriate protein background. Therefore, some of the motif specificities derived from the analyses presented here may not be accurate, depending on the experimental conditions. The purpose of this work was not to study particular lectin specificities, but rather to develop and validate a method for the rapid and unbiased extraction of the information that exists in glycan array data. This method provides a means for researchers to systematically compare the effects of experimental conditions such as lectin concentration, lectin labeling method, glycan density, and wash conditions, in order to define optimal conditions and better distinguish specific from nonspecific interactions. The valency and orientation questions need to be addressed through new glycan array technologies, but the systematic analysis and comparison of the results should be greatly facilitated by motif-based analysis. A new development in glycan array technology that addresses some current limitations is fluidic glycan microarrays (Zhu et al. 2009).

This analysis permits not only the rapid and systematic analysis of lectin specificities, but also global comparisons and searches among lectins and motifs (Figure 5). Such a cataloging of information will be valuable to identify lectins that might bind particular glycan structures or to classify lectins according to similarities in specificities. With the continued use and optimization of glycan arrays, these analyses will further support the integration of glycome-wide and proteome-wide information in biological studies. The database we developed containing the results of these analyses (available through the external links page at www.functionalglycomics.org) is a first step toward that goal.

In conclusion, we demonstrate here the development of a new method for the systematic analysis of glycan array data to identify and define lectin specificities. The method should be broadly useful for the analysis of any type of glycan array format and glycan-binding protein. The rapid, automated analysis of glycan array data could increase the value of the experiments by providing more efficient optimization of experimental conditions, the ability to objectively classify both simple and complex binding specificities, and a means for comparing and searching among multiple experiments. These developments have implications for the more effective use of glycan-binding proteins as analytical reagents and for the improved understanding of the roles of lectins in biology.

Material and methods

Data source

The glycan array data were obtained from the Consortium for Functional Glycomics (CFG) website (www.functionalglycomics.org). Data available as of August 2008 were obtained. Glycan arrays with no fluorescence values above an intensity of 3000 were excluded. The list of plant lectins, animal lectins, and glycan-binding antibodies represented, along with the identifier of the glycan array experiment used for each, is provided in supplementary Tables V–VII, respectively.

Generation of glycan array data

The glycan array experiments were performed by Core D of the Consortium for Functional Glycomics, as described previously (Blixt et al. 2004). A brief description of the experimental process is give here. Synthetic glycans functionalized with a spacer and a terminating NH_2 group were spotted onto NHS-activated microscope slides (Slide H, Schott Nexterion) using a robotic microarrayer. Lectins and glycan-binding antibodies at a concentration of 5–50 μg/mL in a buffer of PBS containing 0.005%–0.5% Tween-20 were incubated on the arrays for 30–60 min. The lectins and antibodies were either tagged with a fluorophore (Alexa Fluor 488, Molecular Probes) or biotin, or they were used unlabeled. If fluorophore-labeled analytes were used, the arrays were washed and immediately scanned for fluorescence using a microarray scanner. Biotinylated analytes were detected with an incubation of streptavidin-FITC, and unlabeled analytes were detected with an appropriate FITC-labeled antibody, followed by washing and scanning. Image analysis software was used to quantify the fluorescence intensities at each glycan spot. The data from six replicate spots were averaged to achieve a final value.

Data analysis

The primary data analysis and calculations were performed with Microsoft Office Excel 2007. The clusters were created using the programs Cluster/Treeview and MultiExperiment Viewer, and the figures were created in Deneba Canvas X.

The intensity segregation method used a calculation of the difference in the fractional representation of each motif between the high-intensity glycans and the low-intensity glycans. That calculation used the formula: $(G_{mb}/G_b) - (G_{mu}/G_u)$, where G_{mb} is the number of glycans in the bound group that have the motif, G_b is the total number of glycans in the bound group, G_{mu} is the number of glycans in the unbound group that have the motif, and G_u is the total number of glycans in the unbound group.

The motif segregation method used a calculation of the statistical difference between the intensities of the glycans containing a particular motif and intensities of the glycans not containing the motif. We used the *P*-value from the Mann–Whitney nonparametric test for that purpose. For the purpose of graphically representing the scores, we log transformed (base 10) the *P*-values. To indicate whether the motif-containing or the non-motif-containing glycans had the higher values, we multiplied the transformed *P*-values by the sign of the *z*-score.

A relational database was created using PostgreSQL to facilitate interrogation of the experimental results. Data were imported into the relational schema as tables that capture the relationships between the motifs, the glycan-binding proteins, and the logged *P*-values calculated from the data. The database was encapsulated by a web-based interface developed using J2EE technologies running on a JBoss application server. This dynamic, database-driven web page allows filtering of the experimental results by glycan motif/lectin combination, lectin category (animal, plant, or antibody), glycan array version, and logged *P*-values within a user-defined specification.

Supplementary data

Supplementary data for this article is available online at http://glycob.oxfordjournals.org/.

Acknowledgements

We thank Dr. James Paulson for constructive review of this work, and Dr. Kyle Furge and Karl Dykema of the VARI Bioinformatics Core and Dr. David Cherba (VARI) for computational support. We thank Dr. Yi-Mi Wu and Anusha Sunkara (VARI) for technical assistance in assembling the data.

Funding

The Van Andel Research Institute and the National Cancer Institute (R33 CA122890, to B.B.H.).

Conflict of interest statement

None declared.

Abbreviations

CFG, Consortium for Functional Glycomics; Fuc, fucose; Gal, galactose; GalNAc, *N*-acetylgalactosamine; Glc, glucose; GlcNAc, *N*-acetylglucosamine; Man, mannose.

References

Angeloni S, Ridet JL, Kusy N, Gao H, Crevoisier F, Guinchard S, Kochhar S, Sigrist H, Sprenger N. 2005. Glycoprofiling with micro-arrays of glycoconjugates and lectins. *Glycobiology.* 15:31–41.

Blixt O, Head S, Mondala T, Scanlan C, Huflejt ME, Alvarez R, Bryan MC, Fazio F, Calarese D, Stevens J et al. 2004. Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc Natl Acad Sci USA.* 101:17033–17038.

Chen S, LaRoche T, Hamelinck D, Bergsma D, Brenner D, Simeone D, Brand RE, Haab BB. 2007. Multiplexed analysis of glycan variation on native proteins captured by antibody microarrays. *Nat Methods.* 4:437–444.

Culf AS, Cuperlovic-Culf M, Ouellette RJ. 2006. Carbohydrate microarrays: Survey of fabrication techniques. *Omics.* 10:289–310.

Dennis JW, Granovsky M, Warren CE. 1999. Protein glycosylation in development and disease. *Bioessays.* 21:412–421.

Dube DH, Bertozzi CR. 2005. Glycans in cancer and inflammation—Potential for therapeutics and diagnostics. *Nat Rev Drug Discov.* 4:477–488.

Freeze HH. 2006. Genetic defects in the human glycome. *Nat Rev Genet.* 7:537–551.

Freeze HH, Aebi M. 2005. Altered glycan structures: The molecular basis of congenital disorders of glycosylation. *Curr Opin Struct Biol.* 15:490–498.

Fuster MM, Esko JD. 2005. The sweet and sour of cancer: Glycans as novel therapeutic targets. *Nat Rev Cancer.* 5:526–542.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science.* 286:531–537.

Hirabayashi J. 2004. Lectin-based structural glycomics: Glycoproteomics and glycan profiling. *Glycoconj J.* 21:35–40.

Hirabayashi J, Arata Y, Kasai K. 2003. Frontal affinity chromatography as a tool for elucidation of sugar recognition properties of lectins. *Methods Enzymol.* 362:353–368.

Hirabayashi J, Hashidate T, Arata Y, Nishi N, Nakamura T, Hirashima M, Urashima T, Oka T, Futai M, Muller WE et al. 2002. Oligosaccharide specificity of galectins: A search by frontal affinity chromatography. *Biochim Biophys Acta.* 1572:232–254.

Kuno A, Uchiyama N, Koseki-Kuno S, Ebe Y, Takashima S, Yamada M, Hirabayashi J. 2005. Evanescent-field fluorescence-assisted lectin microarray: A new strategy for glycan profiling. *Nat Methods.* 2:851–856.

Lau KS, Dennis JW. 2008. *N*-Glycans in cancer progression. *Glycobiology.* 18:750–760.

Li C, Simeone D, Brenner D, Anderson MA, Shedden K, Ruffin MT, Lubman DM. 2009. Pancreatic cancer serum detection using a lectin/glyco-antibody array method. *J Proteome Res.* 8:483–492.

Manimala JC, Roach TA, Li Z, Gildersleeve JC. 2007. High-throughput carbohydrate microarray profiling of 27 antibodies demonstrates widespread specificity problems. *Glycobiology.* 17:17C–23C.

Osako M, Yonezawa S, Siddiki B, Huang J, Ho JJ, Kim YS, Sato E. 1993. Immunohistochemical study of mucin carbohydrates and core proteins in human pancreatic tumors. *Cancer.* 71:2191–2199.

Patwa TH, Zhao J, Anderson MA, Simeone DM, Lubman DM. 2006. Screening of glycosylation patterns in serum using natural glycoprotein microarrays and multi-lectin fluorescence detection. *Anal Chem.* 78:6411–6421.

Pilobello KT, Krishnamoorthy L, Slawek D, Mahal LK. 2005. Development of a lectin microarray for the rapid analysis of protein glycopatterns. *Chembiochem.* 6:985–989.

Satomura Y, Sawabu N, Takemori Y, Ohta H, Watanabe H, Okai T, Watanabe K, Matsuno H, Konishi F. 1991. Expression of various sialylated carbohydrate antigens in malignant and nonmalignant pancreatic tissues. *Pancreas.* 6:448–458.

Sharon N. 2007. Lectins: Carbohydrate-specific reagents and biological recognition molecules. *J Biol Chem.* 282:2753–2764.

Shimizu K, Katoh H, Yamashita F, Tanaka M, Tanikawa K, Taketa K, Satomura S, Matsuura S. 1996. Comparison of carbohydrate structures of serum alpha-fetoprotein by sequential glycosidase digestion and lectin affinity electrophoresis. *Clin Chim Acta.* 254:23–40.

Tateno H, Nakamura-Tsuruta S, Hirabayashi J. 2007. Frontal affinity chromatography: Sugar–protein interactions. *Nat Protoc.* 2:2529–2537.

Tateno H, Nakamura-Tsuruta S, Hirabayashi J. 2009. Comparative analysis of core-fucose-binding lectins from *Lens culinaris* and *Pisum sativum* using frontal affinity chromatography. *Glycobiology.* 19:527–536.

Varki A, Cummings R, Esko J, Freeze H, Hart G, Marth J. 1999. *Essentials of Glycobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Wang D. 2003. Carbohydrate microarrays. *Proteomics* 3:2167–2175.

Wearne KA, Winter HC, O'Shea K, Goldstein IJ. 2006. Use of lectins for probing differentiated human embryonic stem cells for carbohydrates. *Glycobiology*. 16:981–990.

Yue T, Goldstein IJ, Hollingsworth MA, Kaul K, Brand RE, Haab BB. 2009. The prevalence and nature of glycan alterations on specific proteins in pancreatic cancer patients revealed using antibody-lectin sandwich arrays. *Mol Cell Proteomics.* 8:1697–1707.

Zhu XY, Holtz B, Wang Y, Wang LX, Orndorff PE, Guo A. 2009. Quantitative glycomics from fluidic glycan microarrays. *J Am Chem Soc.* 131:13646–13650.