

CarboGrove: a resource of glycan-binding specificities through analyzed glycan-array datasets from all platforms

Zachary L Klamer¹ , Chelsea M Harris², Jonathan M Beirne², Jessica E Kelly², Jian Zhang², Brian B Haab^{1,*}

¹Department of Cancer and Cell Biology, Van Andel Institute, 333 Bostwick Ave NE, Grand Rapids, MI 49503, United States, ²Z Biotech, Aurora, CO, United States

*Corresponding author: Van Andel Institute, 333 Bostwick Ave NE, Grand Rapids, MI 49504, United States. Email: brian.haab@vai.org

Glycan arrays continue to be the primary resource for determining the glycan-binding specificity of proteins. The volume and diversity of glycan-array data are increasing, but no common method and resource exist to analyze, integrate, and use the available data. To meet this need, we developed a resource of analyzed glycan-array data called CarboGrove. Using the ability to process and interpret data from any type of glycan array, we populated the database with the results from 35 types of glycan arrays, 13 glycan families, 5 experimental methods, and 19 laboratories or companies. In meta-analyses of glycan-binding proteins, we observed glycan-binding specificities that were not uncovered from single sources. In addition, we confirmed the ability to efficiently optimize selections of glycan-binding proteins to be used in experiments for discriminating between closely related motifs. Through descriptive reports and a programmatically accessible Application Programming Interface, CarboGrove yields unprecedented access to the wealth of glycan-array data being produced and powerful capabilities for both experimentalists and bioinformaticians.

Key words: binding specificity; database; glycan-binding protein; lectin; microarray.

Introduction

Glycan arrays are being produced and used by more labs than ever before. After the first reports of glycan arrays in 2004 (Blixt et al. 2004; Bryan et al. 2004), just a handful of laboratories worked on the technology for about the next decade. The main provider of the technology was the Consortium for Functional Glycomics (CFG). The CFG array, which used the planar array method that had been established for DNA and protein arrays, contained glycans that represented a broad survey of the known, important motifs in mammalian biology. The significance of the CFG resource was that it provided access to researchers who could not produce arrays themselves, primarily by reason of the cost and difficulty of synthesizing glycans. It was the large number of experiments performed on this platform that established the value of glycan-array technology and stimulated further developments in the field, including in experimental methods, bioinformatics tools, and methods of glycan synthesis. The CFG data were the sole data source for multiple bioinformatics efforts in the analysis of glycan array data (Porter et al. 2010; Cholleti et al. 2012; Cao et al. 2019; Coff et al. 2020).

But no single array could meet the needs of every study. The diversity and number of structures present among various classes of glycans and organism types are too great even for the largest array. This situation drove researchers to develop arrays with content organized around specific fields of research. For example, plant biologists and microbial biologists each developed glycan arrays relevant to their fields (Ruprecht et al. 2017; Geissner et al. 2019), and researchers studying sialic acids developed arrays with a wide range of variants of that feature (Song et al. 2011). Other features of specialized content include glycosaminoglycans

(Horton et al. 2020; Chopra et al. 2021), glycopeptides (Hinou et al. 2019; Mende et al. 2020; Nason et al. 2021), and human milk oligosaccharides (Prudden et al. 2017). Improved synthetic strategies and automated synthesizers helped address the major hurdle of glycan production (Zhang et al. 2018; Li, Liu, et al. 2019b), in combination with focused production around specific types of glycans. Furthermore, researchers developed experimental alternatives to the planar array. The novel technologies—methods involving mass spectrometry (Kitov et al. 2019) or bacteriophage display (Sojitra et al. 2021), for example—provide complementary information and capabilities to the planar array and allow dispersion of the methods to a greater number of laboratories.

All of these developments have resulted in an expanding variety of glycan-array data available for study. Bioinformatics methods that could capture and use all available glycan-array data, regardless of source and content, could serve many purposes, from learning more about the specificity of a particular protein, to finding lectins with a pre-defined specificity, to larger-scale, integrative studies in glycobiology. Such analyses cannot be done manually, given the complexity of glycans and protein–glycan interactions, as well as the complexity of integrating information over many data points from the array. The data from the various sources must be analyzed and interpreted with a common system.

Several resources currently provide glycan-array data in either raw or analyzed form: CFG, Lectin Frontier Database (LFDB; Hirabayashi et al. 2015), Multiple Carbohydrate Alignment with Weights Database (MCAW-DB; Hosoda et al. 2018), and Glycan Microarray Database (GlyMDB; Cao et al. 2019). These resources represent valuable advances in the field, but they have limited value as a general resource for

Received: January 12, 2022. Revised: March 24, 2022. Accepted: March 24, 2022

© The Author(s) 2022. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

non-bioinformaticians. One limitation is that each provides data for only a single type of array, either CFG (CFG, MCAW-DB, and GlyMDB) or frontal-affinity chromatography (LFDB). Further, there are resources which provide powerful visualization tools, such as the GLAD (GLycan Array Dashboard) employed by GlyMDB, but limited or incomplete interpretation of the data. For example, MCAW-DB gives an alignment of top-binding glycans, which gives clues about features associated with binding, but it does not provide the context of their binding strength relative to others, which is necessary to achieve a complete picture. In general, the discernment of the specificity of a glycan-binding protein requires significant, additional analysis on the part of the researcher. The need for algorithms to discern the complex, fine specificities of glycan-binding proteins is clear, given the sensitivity of binding to minor differences in glycans such as the position of the epitope on *N*-glycan branches (Li, Guan, et al. 2019a; Wang et al. 2021).

Given software to reliably interpret data from any platform, a resource could be built that provides common access to glycan-array information across the many sources that are now available. We recently introduced the MotifFinder software (Klamer et al. 2017; Klamer and Haab 2021) to meet the analysis need. We demonstrated earlier that the algorithm delivers a detailed and accurate analysis of the specificity of a glycan-binding protein and that it can perform the analyses through the integration of data from distinct platforms. These previous developments suggested an approach to unify the analysis and usage of glycan array data. In the present work, we explored whether a database system driven by the MotifFinder engine could meet the need for a unified glycan-array resource.

Results

Achieving common data processing across platforms and array types

The available glycan-array data cover several approaches to detection and quantification (Fig. 1A). This diversity represents a challenge when collating data into a common platform, but it also provides complementary information from the strengths and limitations of each platform. The planar array uses robust methods that were established for DNA arrays and thus has been a workhorse for many labs, but it has the limitation that nonspecific or reduced binding can occur from the linker (Grant et al. 2014), the surface (Li et al. 2021), or the tagging of the glycan-binding protein (Kitova et al. 2019). Most embodiments do not account for glycan density or kinetics. The newer technologies provide solution-phase kinetics (Kitov et al. 2019; Sojitra et al. 2021), incorporation of density as a controllable parameter (Mende et al. 2020), or display on a cell-surface context (Büll et al. 2021), but they require specialized methods or equipment.

The available data also contain a great diversity of glycans (Fig. 1B). The CFG array was heavily weighted toward mammalian glycans, but technology developers have branched out into microbial, plant cell wall, and other nonmammalian glycans. In addition, improved synthesis technologies have resulted in glycans with increased complexity and a broader variety of monosaccharides, as well as a variety of glycosaminoglycans. These developments have resulted

in an increased frequency in reports on new glycan arrays (Fig. 1C, Supplementary Table I).

We sought to develop a system that provides a common mode of analysis for all available data. To account for the diversity in glycans, we utilized MotifFinder's glycan parser that translates text representations of all types of monosaccharide names and their connections in standard CFG-like notation to graph structures used by the program. To enable the processing of data from any type of array or platform, we developed the analysis algorithm to be independent of scale or range but require only a quantitative value corresponding to each glycan in the array positively associated with binding.

The various glycan-array data could then be processed in our MotifFinder algorithm for identifying the motifs—patterns within glycans—that best describe the specificity of the glycan-binding protein applied to the array. The algorithm uses data from multiple concentrations of the protein, if available, to give more accurate results than possible from one concentration (Klamer and Haab 2021) (Fig. 2A). The family of motifs that defines the specificity (referred to as the model) is arranged into 2 types: the primary motifs, which represent distinct structural categories, and the fine-specificity motifs, which represent gradations in binding within the primary motifs (Fig. 2B). The model is visualized in various ways to assist user interpretation (Fig. 2C). The consistent output across all datasets, regardless of platform or type of glycans, is a critical component of enabling cross-dataset comparisons and searches.

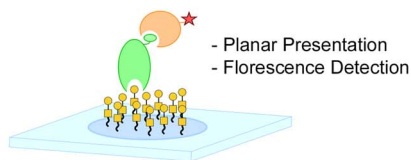
We tested the ability of the algorithm to process and organize glycan-array data from 35 different types of arrays, 13 different glycan families, 5 experimental methods, and 19 laboratories or companies (Fig. 3A, Supplementary Table II). These included publicly available data as well as unpublished data (Supplementary Tables II and III). The number of contributions from each provider ranged from 1 to 541 datasets, for a total of 1,125 datasets (Supplementary Table IV). MotifFinder was able to produce a model for each of the glycan-binding proteins with only minor adjustments in formatting required for some datasets.

We then assembled the models into a relational database called CarboGrove (Fig. 3A). This collection promises to cover a much broader range of glycans and the glycan-binding proteins than any single resource. To evaluate the scope of the database, we defined 118 different motifs from 11 families based on a set obtained from the GlyGen resource (York et al. 2019) with the addition of motifs covering the major core types (Supplementary Tables V and VI). This list is not exhaustive but provides an initial, unbiased survey of the breadth of the database. An analysis of the glycans on the arrays showed that all motifs were represented on at least one array and that some were on nearly every array (Fig. 3B). The arrays had a broader range of inclusion of motifs, ranging from only 1 to 112, reflecting the variation in purposes of the arrays.

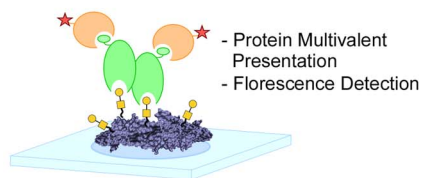
To determine whether the glycan-binding proteins in the database cover a broad range of specificities, we used the model for each protein to predict binding to each of 1,803 glycans that spanned all arrays. From the resulting values, we assessed binding to each of the 118 motifs using a motif score (Klamer and Haab 2021). All but 19 of the 118 motifs are bound (motif score > 2) by 4 or more glycan-binding proteins (Fig. 3C, Supplementary Table V). The 19 motifs not bound by any proteins represent less-common features such

A Array Technology

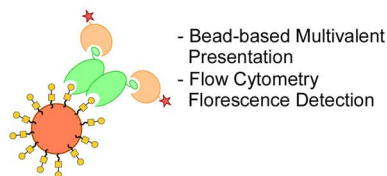
Traditional Array



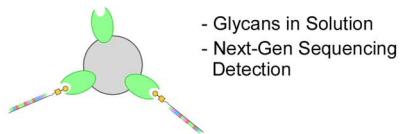
Neoglycoprotein



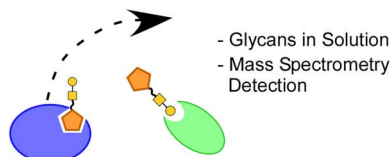
Glycan Bead



NGGM



CUPRA

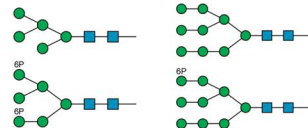


LiGA

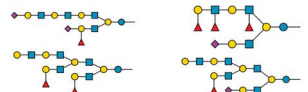


B Array Content

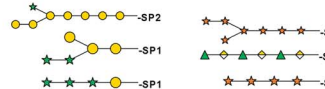
Man 6-Phosphate



Human Milk Oligosaccharides



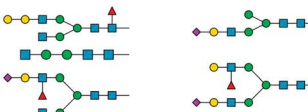
Plant Cell Wall Array



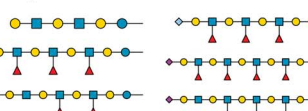
Microbial Array



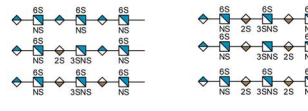
Asymmetric N-Glycan Array



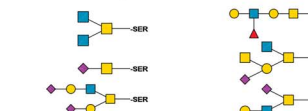
PolyLacNAc



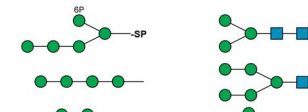
Heparan Sulfate



Chemoenzymatic O-GalNAc



Oligomannose



C Glycan Array Timeline

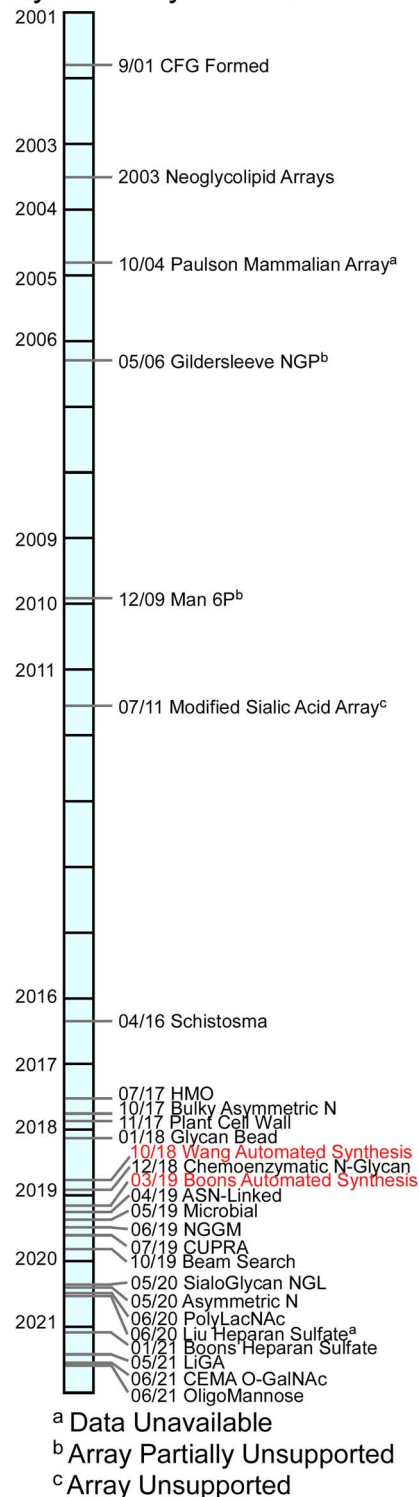


Fig. 1. Diversity of glycan-array technology and content. A) Several technologies in addition to the planar array are now used to probe glycan arrays. The arrays differ in their modes of glycan or protein presentation and in methods of quantification. B) The sets of glycans contained in the arrays represent diverse types of structures and organisms. C) The rate of development of new arrays has increased since 2016, punctuated by significant advances in glycan-synthesis technology (red text). Monosaccharide symbols follow the SNFG (Symbol Nomenclature for Glycans) system (PMID 26543186, *Glycobiology* 25: 1323–1324, 2015) details at NCBI.

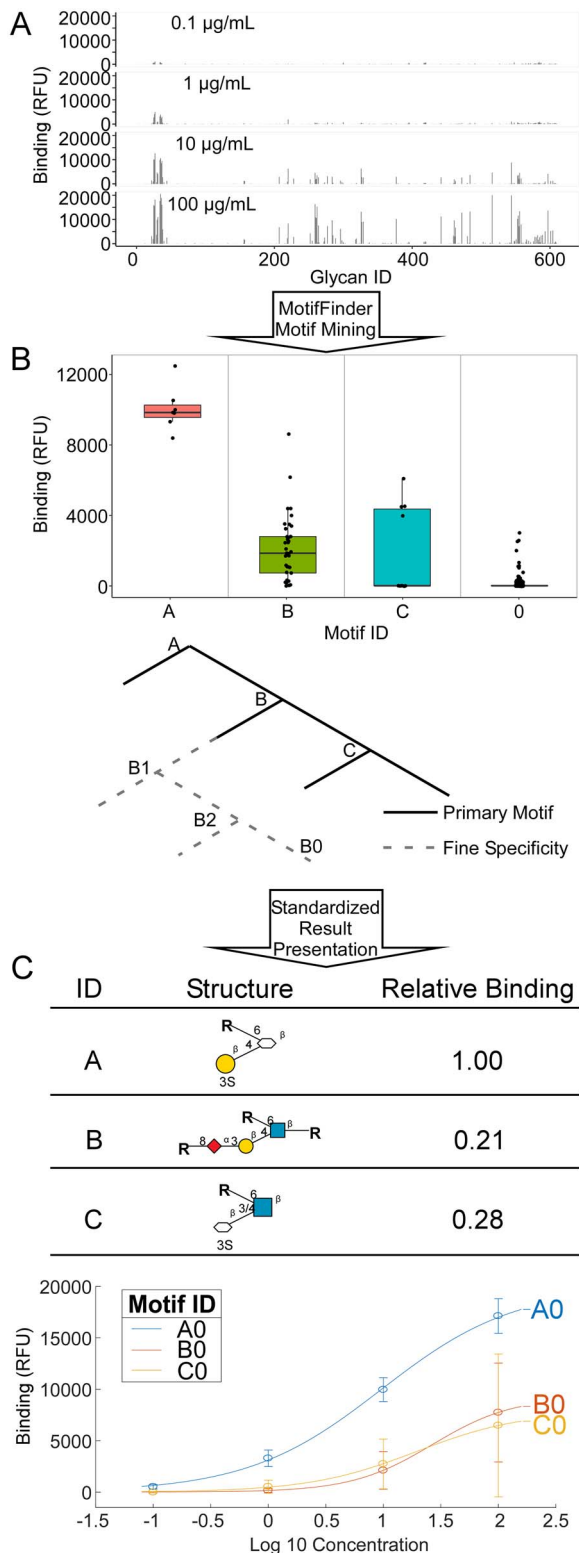


Fig. 2. Standardized analysis and output. A) MotifFinder analyzes glycan-array data from multiple incubation concentrations, where available, of a given glycan-binding protein. B) The program identifies the family of motifs that represents the specificity of the protein, organized into primary motifs and the subtrees of fine-specificity motifs. C) Among the several visualizations in the output are tabular descriptions and binding curves for each motif. Monosaccharide symbols follow the SNFG (Symbol Nomenclature for Glycans) system (PMID 26543186, *Glycobiology* 25: 1323–1324, 2015) details at NCBI.

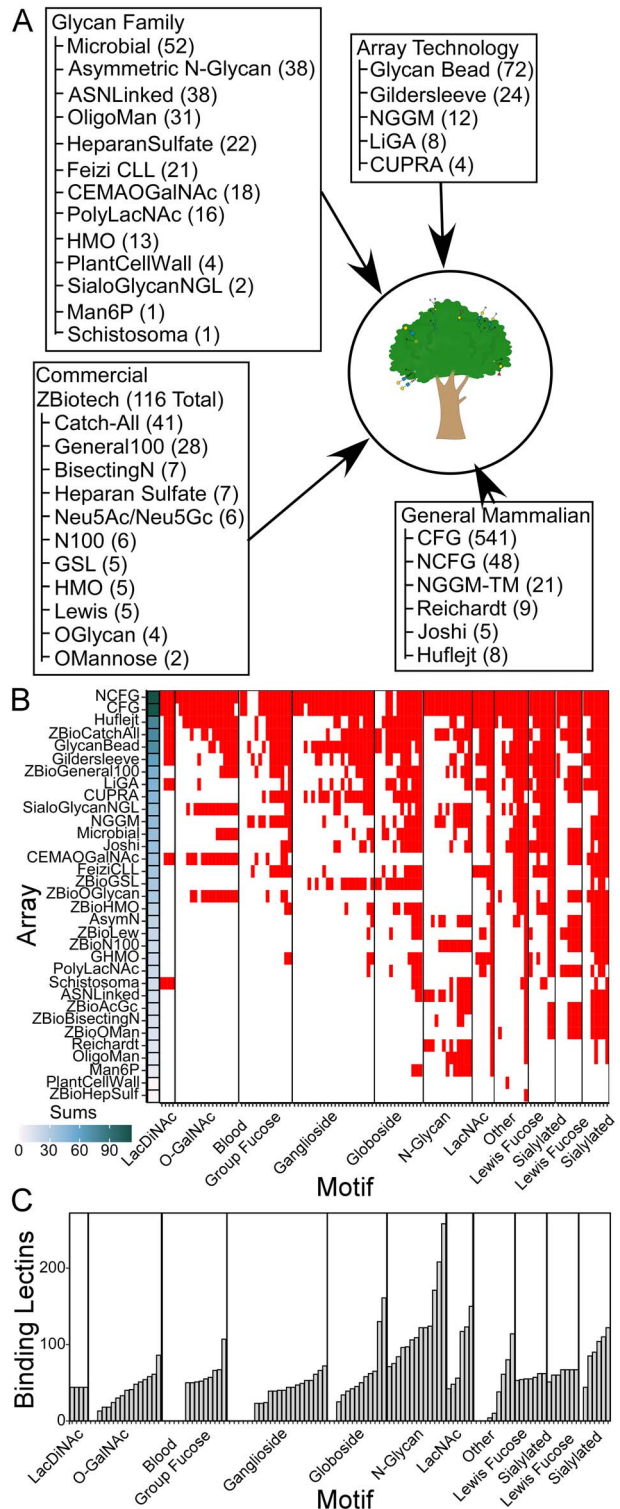


Fig. 3. Breadth of representation of array types of glycan-binding specificities. A) The collection covers a wide variety of glycan families, array providers, and technologies. B) Motif coverage across the arrays. Nearly all motifs are represented on at least 1 array. C) Motif coverage of the glycan-binding proteins.

as type 3 A antigen. Very common motifs such as N-glycan and biantennary N-glycan are bound by many glycan-binding proteins: 258 and 208, respectively. These differences reflect both the prevalence of the motif in biology and the amount of research centered on the motif.

Accessing and analyzing glycan arrays across platforms

The collection of analyzed data potentially offers access to detailed, accurate information about the specificity of any given glycan-binding protein. We sought to enable such searches through a system of matching user-specified terms with all relevant datasets. This task involved accounting for variability in the conventions in common names, abbreviations, and the use of the terms lectin and agglutinin. To address this difficulty, we included multiple aliases for each protein and allowed relevant results to be returned even when a search does not match the primary name in the database.

We tested the search and analysis capabilities using the lectin SNA (*Sambucus nigra* agglutinin). The primary specificity of SNA, α 2,6-linked sialic acids, is well known, but the fine specificities are not well understood owing to limited variety in glycans containing α 2,6-linked sialic acid on the arrays and the complexity in the analysis. A search for SNA returned 36 individual datasets from 16 sources (Fig. 4). The datasets had widely varying ranges of dataset noise, as assessed by the reliability score (Fig. 4A), but nearly every data confirmed the canonical specificity of SNA. The top motifs for each array also revealed complementary information. The ASN-linked and AsymmetricN arrays, which focus on N-linked glycans, identified a preference for the tri/tetra-antennary presentation (motifs A1 and A2, ASN-linked array) over the biantennary presentation (motif A0, ASN-Linked array), as well as a preference for the 3' mannose branch (motif A0, AsymmetricN array) over the 6' mannose branch or unbranched presentations (motif B0, AsymmetricN array). The CFG array, which has the greatest diversity in the α 2,6-sialyl-LacNAc motif, identified preferential binding on extended N-linked glycans over O-linked glycans (motifs A0 and A4). Some arrays had limited variation in the α 2,6-sialyl epitope and consequently produced ambiguous, incomplete motifs, such as motifs B0 and D0 on the Gildersleeve array and A0 and the NGGM-TM, NGGM, and Chemoenzymatic Modular Assembly O-GalNAc arrays. Arrays with a low reliability score, such as the LiGA and GlycanBead arrays, generated various weak motifs that are inconsistent with results from other arrays.

We also tested this functionality on the lectin wheat germ agglutinin, which is widely used but has a specificity that is poorly understood. Part of the challenge is its breadth in specificity, binding nearly half of the glycans on the CFG array (273/609 glycans) when applied at high concentrations. The search gave results from 12 different arrays (Supplementary Fig. 1). A comparative analysis identified both known and novel features, such as highest binding to 6'-linked terminal *N*-acetyl-glucosamine (GlcNAc) and 3'-linked Neu5Ac; binding to both GlcNAc and *N*-acetyl galactosamine (GalNAc) in other terminal linkages, provided the 3' carbon is unsubstituted; and potential binding to the heparan sulfate motif GlcNAc1-4GlcA. The novel observations would require experimental confirmation, but they are structurally plausible and demonstrate findings that are made possible through broad analyses of glycan-array data.

Selecting glycan-binding proteins for experimental design

A companion capability is to select a motif from the motif-sort options and search for glycan-binding proteins that bind the motif. We selected 3 motifs for a test of this function: N-glycan

core fucose, Lewis X, and type-2 blood group B. These motifs have the common feature of fucose, but they differ in the fucose linkage: either to the 6', 3', or 2' carbon of the adjoining monosaccharide. A hierarchical cluster of all the models in the database and the set of motifs defined above indicated that the search motifs are bound by separate groups of proteins (Fig. 5A). The top 10 glycan-binding proteins for each motif confirmed that each motif returned a unique set of proteins known to bind the motif (Table 1). An assessment of the top motifs bound by the glycan-binding proteins showed that each protein is a specific binder of the search motif (Fig. 5B).

A closer analysis showed the value of an unbiased search. For example, the top hits for motif "N-Glycan a6 Core Fucose" did not include the lectin commonly used for this motif, *Lens Culinaris* Agglutinin (LCA), because LCA bound many *N*-glycans without alpha-6 fucose (as shown in the CarboGrove model, not shown). The Lewis X and blood group B searches likewise returned results that were unexpected but potentially useful, such as the strong binding to Lewis X of both the anti-Lewis X antibody and the lectins LTA, CCL2, and AAL.

This functionality suggested an additional opportunity in experimental design. Searches such as demonstrated above could be modified to optimally select glycan-binding proteins for an experiment, such as to distinguish between motifs in a biological sample that are difficult to distinguish by mass spectrometry. We tested this concept for terminal GalNAc, in either the alpha or beta orientation, and terminal GlcNAc, in either the bisecting or outer-arm position. These features are isomers but have important differences in biological function.

We sought to identify a minimal set of lectins (limited to 3–4) that would give optimal distinction between the comparison motifs. First, for each of the 4 terminal features (alpha-GalNAc, beta-GalNAc, outer-arm GlcNAc, and bisecting GlcNAc), we defined glycans containing the feature (Fig. 6A). We also defined negative-control glycans that have the core structures but not the terminal features. Next, we searched the database to identify lectins that bind any of the motifs. We predicted the binding of each lectin to each glycan and assembled the values (Fig. 6A), from which we could search for combinations of lectins that give unique patterns of binding across each of the comparison motifs.

Multiple algorithms are available for maximizing distances between subsets. For demonstration, we used manually guided optimization to arrive at the minimal set of GSL-II, HAA, VVL, and PHA-E (Fig. 6B). The average binding of the lectins to the glycans in the comparison groups showed distinct patterns, corresponding to the differences in the top motifs (Fig. 6B): GSL-II binds non-bisecting terminal GlcNAc; HAA binds terminal alpha-GalNAc; VVL binds LacDiNAc and some lipid-linked glycans; and PHA-E binds bisecting GlcNAc. Thus, starting from the full collection of >700 models, we efficiently reduced to just 4 that provide clear distinctions among the 4 isomeric motifs.

Discussion

The proliferation of glycan-array platforms and data has precipitated a need for a common mode of analyzing, interpreting, and accessing the data. Here we provide a solution via the MotifFinder analysis program and the CarboGrove database. We populated the database with analyzed data from

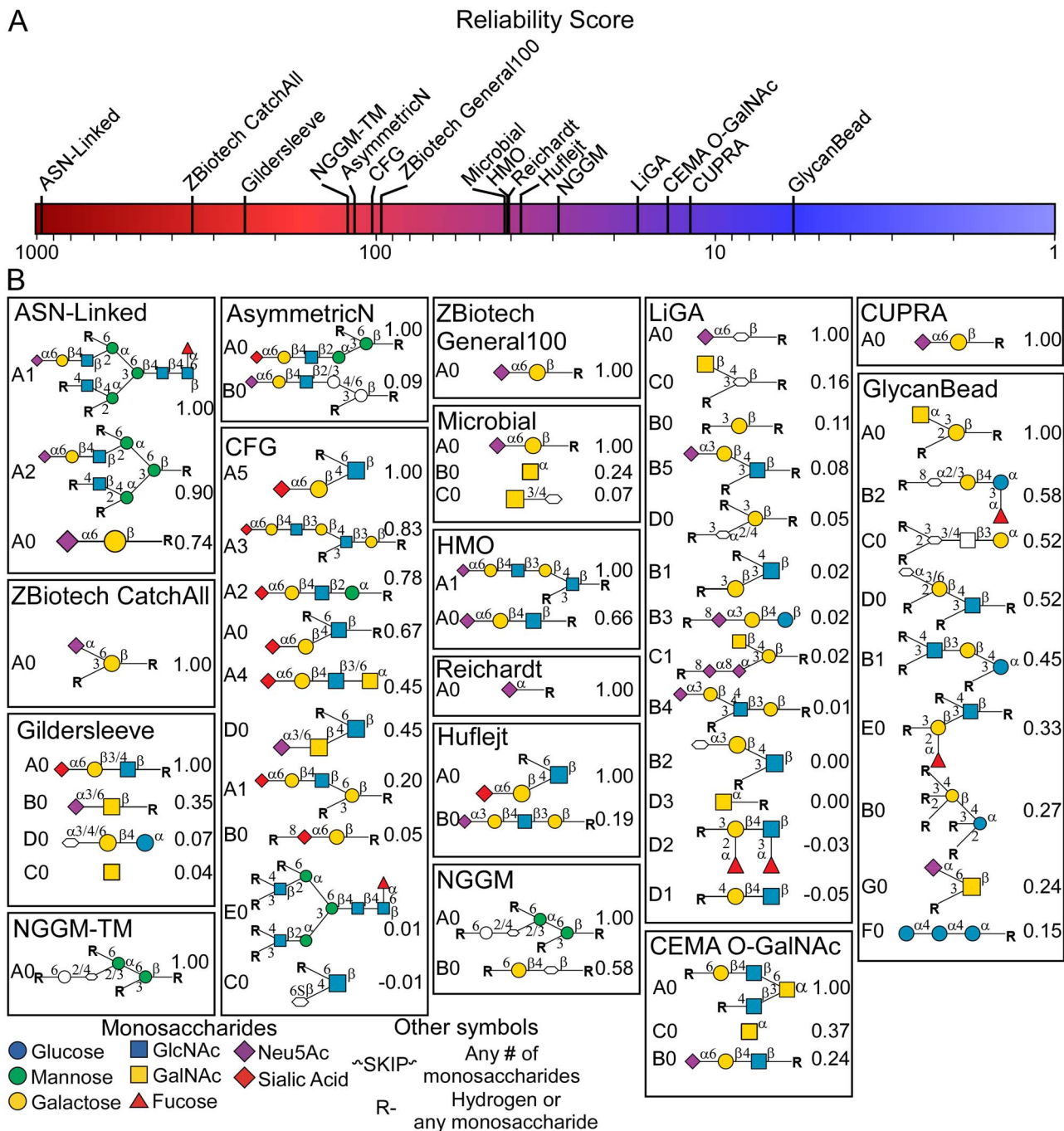


Fig. 4. Comparison across multiple arrays of results for SNA. A) Reliability score for each of the arrays, where higher scores indicate lower dataset noise. B) CarboGrove reported motifs for each array. Within each array, the motifs are ordered from top to bottom by the relative-binding score given next to the ID and graphical representation of the motif. The monosaccharide symbols follow the SNFG (Symbol Nomenclature for Glycans) system (PMID 26543186, *Glycobiology* 25: 1323–1324, 2015) details at NCBI.

35 types of arrays and from multiple suppliers and platforms. Thus, for the first time, researchers can access and use an expansive collection of glycan-array data in an analyzed form. The ability to bring together data from multiple sources is especially important in the case of glycan arrays, where each type of array provides information or experiments that are complementary to the others. In particular, the arrays are complementary in their glycans, the glycan-binding proteins analyzed, and the strengths and weaknesses of their experimental systems. The enabling component of this project was

a software tool for analyzing all types of glycan-array data. Without a common model of analyzing and reporting the data, the assembly and integrated analysis of such diverse data would be prohibitively time-consuming and inaccurate.

The ability to compare and integrate results between separate datasets has advantages for many applications. In the evaluation of SNA, for which the search returned 36 datasets, we observed fine specificities that are distinct from the canonical specificity and that would be missed in the evaluation of only a single platform. The findings were consistent with

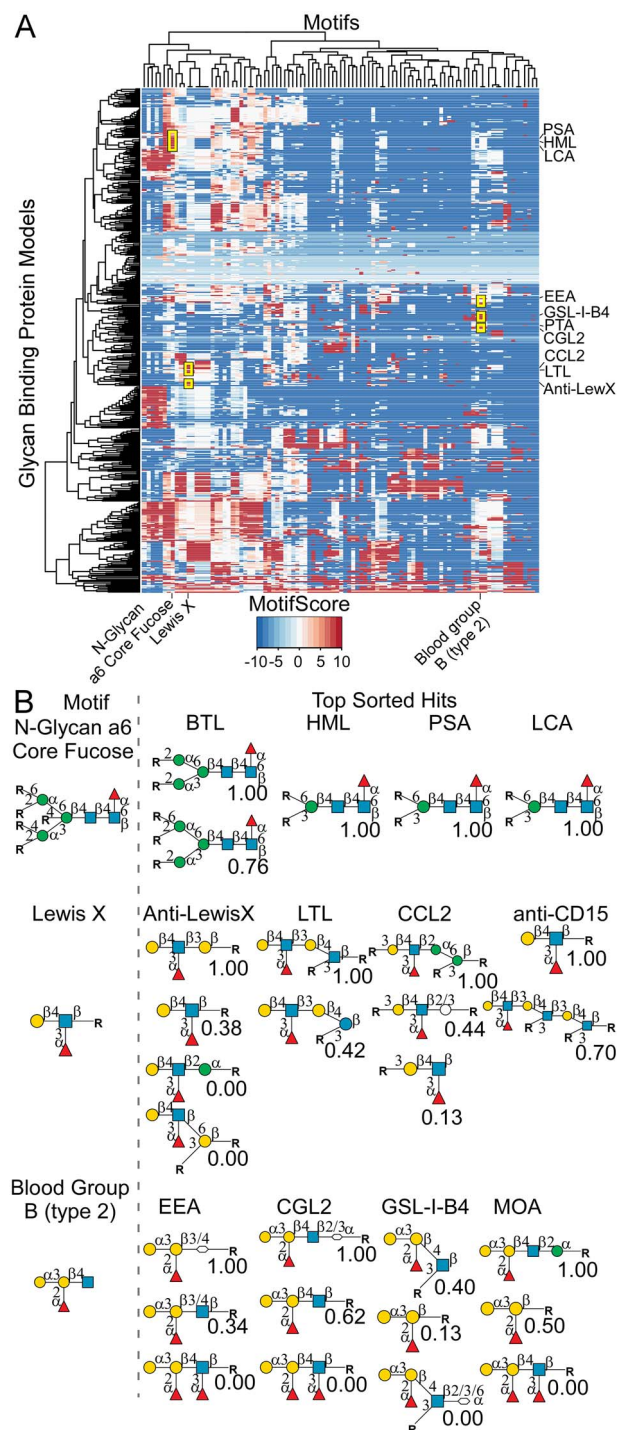


Fig. 5. Searches for glycan-binding proteins that are specific for selected motifs. A) The motif scores for 118 motifs (listed in the columns) and 785 glycan-binding proteins were hierarchically clustered. The search motifs and top hits are labeled. B) For the top 4 glycan-binding proteins from each search, the proteins' top motifs and their relative binding scores are indicated. Monosaccharide symbols follow the SNFG (Symbol Nomenclature for Glycans) system (PMID 26543186, *Glycobiology* 25: 1323–1324, 2015) details at NCBI.

the previous studies of SNA. For example, the MotifFinder identification of preferential binding to the primary epitope on extended N-linked glycans from the 3' mannose branch agrees with the manual analysis (Li, Guan, et al. 2019a). But the ability to conveniently supplement the study with

comparable findings from other arrays, such as the preference for tri/tetra-antennary N-glycans over linear presentations (Fig. 4), is unique.

These analyses also expand on previous methods to compare between platforms. Previous studies to compare between conditions or platforms generally used manual analyses, such as comparisons between arrays that focused on sialic acids (Padler-Karavani et al. 2012) or between experimental conditions (Temme et al. 2019; Li, Guan, et al. 2019a). An algorithm-based approach was introduced in a well-designed study to compare results between separate array platforms (Wang et al. 2014). The authors used a universal thresholding method to identify differences between platforms that correlated with experimental differences. The method did not, however, capture both weak and strong binding for all arrays and did not account for fact that no single concentration was relevant in comparing results between arrays. In contrast, the algorithm used in the present work is sensitive to weakly bound glycans and allows comparisons among arrays even where concentrations are not optimally matched. Furthermore, the software has the unique support of combined analyses of multiple datasets, previously demonstrated for integrating data from multiple concentrations of a lectin or from multiple platforms (Klamer and Haab 2021).

This study also highlighted the need for a standardized system of reporting the details of glycan array experiments. The MIRAGE guidelines for reporting glycan-array metadata (Supplementary Table II) help to improve accessibility of metadata and the ability to compare results (Liu et al. 2017). Tools designed to standardize the design, processing, and storage of glycan array data like the in-development CarbArrayArt and GlyGen glycan array database will further improve the accessibility of glycan array metadata (Mehta et al. 2020).

Besides studying the specificities of glycan-binding proteins, a major function of CarboGrove is to identify glycan-binding proteins that have user-specified binding traits. This function will be important for both non-experts and experts in glycobiology. Searches across lesser-known platforms and motifs can be impractical, and the specificities of some glycan-binding proteins are not generally known or are misunderstood. But even more valuable could be the ability to select optimal sets of lectins for experiments. The use of lectins to detect or quantify glycan structures is very common, for example in methods such as immunofluorescence, western blot, cell staining, in vivo imaging, and others, but in many cases, the experiments do not employ the optimal lectins or are inaccurately interpreted. Using CarboGrove, a researcher could perform searches to identify a limited number of glycan-binding proteins that target the motifs that are relevant to the biological study. MotifFinder could predict binding to a set of relevant glycans and select the proteins most useful for analysis. Bioinformaticians could support this work by developing tools employing additional approaches to optimize experiments that use glycan-binding proteins.

The current study and resource have several limitations. The analysis algorithm does not account for certain features that could influence binding, such as the method of attaching the glycan, the density of the glycan, or the nature of a polypeptide backbone if present. A wide variety of experimental conditions also can influence apparent binding, including buffer, array substrate, and detection method (Kitova et al. 2019). The resource currently does not house much of the

Table 1. Top 10 glycan-binding proteins for each search motif.

Motif	Rank	Lectin	Name	Canonical motif	Source	Array
N-Glycan Core a6 Fucose	1	BTL	Bryothamnion Triquetrum Lectin		Investigator	CFG
	2	HML	<i>Hypnea Musciformis</i> Lectin		Investigator	CFG
	3	PSA	<i>Pisum Sativum</i> Agglutinin	Core fucose	EY Labs	CFG
	4	PSA	<i>Pisum Sativum</i> Agglutinin	Core fucose	Vector	CFG
	5	rBTL	Bryothamnion Triquetrum Lectin		Investigator	CFG
	6	PSA	<i>Pisum Sativum</i> Agglutinin	Core fucose	Vector	ASNLinked
	7	LCA	<i>Lens Culinaris</i> Agglutinin	Core fucose	Vector	ASNLinked
	8	LCA	<i>Lens Culinaris</i> Agglutinin	Core fucose	Vector	CFG
	9	LCA	<i>Lens Culinaris</i> Agglutinin	Core fucose	Vector	CFG,NCFG
	10	PSA	<i>Pisum Sativum</i> Agglutinin	Core fucose	Vector	ZBiotech
Lewis X	1	Anti-LewX	Clone 28 Anti-Lewis X Antibody	Lewis X	Investigator	CFG
	2	LTL	<i>Lotus Tetragonolobus</i> Lectin	Fucose	Vector	PolyLacNAc
	3	CCL2	Coprinopsis Cinerea Lectin 2	Fucose alpha1,3	Investigator	CFG
	4	anti-CD15	anti-CD15	Lewis X	Investigator	PolyLacNAc
	5	AAL	Aleuria Aurantia Lectin	Fucose	Vector	PolyLacNAc
	6	CCL2	Coprinopsis Cinerea Lectin 2	Fucose alpha1,3	Investigator	CFG
	7	AAL	Aleuria Aurantia Lectin	Fucose	Vector	FeiziCLL
	8	LTL	<i>Lotus Tetragonolobus</i> Lectin	Fucose	Vector	Zbiotech
	9	AAL	Aleuria Aurantia Lectin	Fucose	Vector	CEMAOGalNAc
	10	AAL	Aleuria Aurantia Lectin	Fucose	Vector	AsymN
Blood Group B (type 2)	1	EEA	<i>Euonymus Europaeus</i> Agglutinin	Blood group B	EY Labs	CFG
	2	EEA	<i>Euonymus Europaeus</i> Agglutinin	Blood group B	Vector	GlycanBead
	3	CGL2	Coprinopsis Cinerea Galectin 2	Fucose alpha1,2	Investigator	CFG
	4	GSL-I-B4	Griffonia Simplicifolia Lectin 1, B4	Galactose	Vector	CFG
	5	MOA	Marasmius Oreades Agglutinin	Alpha-galactose	EY Labs	CFG
	6	PTA	<i>Psophocarpus Tetragonolobus</i> Agglutinin	Blood groups	EY Labs	CFG
	7	GSL-I-B4	Griffonia Simplicifolia Lectin 1, B4	Galactose	Vector	CFG,NCFG
	8	EEA	<i>Euonymus Europaeus</i> Agglutinin	Blood group B	Vector	CFG
	9	PA-IL	<i>Pseudomonas Aeruginosa</i> Lectin 1	Galactose	Sigma	CFG
	10	GSL-I-B4	Griffonia Simplicifolia Lectin 1, B4	Galactose	Vector	Zbiotech

metadata associated with the experiments, from which one could explore factors that influence binding. The metadata exists in a great variety of completeness and form, so the inclusion of all available information in the database is a technical hurdle. Currently, the database includes references to the original sources of the data, through which researchers locate complete metadata if needed. While the database does house all the data used in the analyses performed here (accessible via the Application Programming Interface), the database is not intended to be a resource for data deposition, but conceivably such a development could be useful. Finally, enabling support for user-defined motifs for database searches could expand the specificity-finding capabilities of the resource. This addition requires additional developments in the motif-building tools and would dramatically increase the computational overhead of the database and thereby database operating costs, one of the limiting factors in bioinformatics-resource lifecycles.

Another important goal is to provide connections between resources providing complementary information, for example regarding the glycans, the motifs, the GBPs, or the platforms would be valuable. For bioinformatics developers, such connectivity is already available through the Application Programming Interface (API) in CarboGrove. For general users, links to the UniLectin database via the stable UniProt accession numbers of the lectin are currently provided, and links to addition information, such as through GlyTouCan or GlyGen, will be continuously added in updates (Bonnardel et al. 2019;

York et al. 2019; Consortium et al. 2020; Fujita et al. 2020). Many additional developments in bioinformatics could be stimulated by this resource. For example, bioinformaticians could use the resource to explore relationships among families of lectins and antibodies in association with genetic or organismal information. The motifs from MotifFinder could be used as the connection to a wide range of data on sequence, biosynthesis, and other information, such as are accessible through the GlyGen resource (York et al. 2019) and other databases. A resource that appeared after submission of this work is a machine learning annotation of lectin specificities, which could provide additional, complementary information (Bojar et al. 2022).

The pace of introduction of new arrays and glycans is clearly quickening. Versatile systems of attachment to surfaces using both covalent and noncovalent deposition (Li et al. 2021) could make array production easier for nonspecialist labs. Others have displayed glycans on the surface of bacteriophage (Sojitra et al. 2021), produced various glycans through sequential knockdowns of genes in glycan biosynthesis (Büll et al. 2021), and tuned the density of glycans spotted on chip through efficient methods of producing glycopeptide arrays (Mende et al. 2020). The MotifFinder platform supports updating to allow for additional factors to be explored. Our ongoing work involves support for investigating the influence of peptide backbone, glycan density, and linker type, as well as experimental factors that have been shown to introduce variability (Temme et al. 2019). Thus, we present a system to

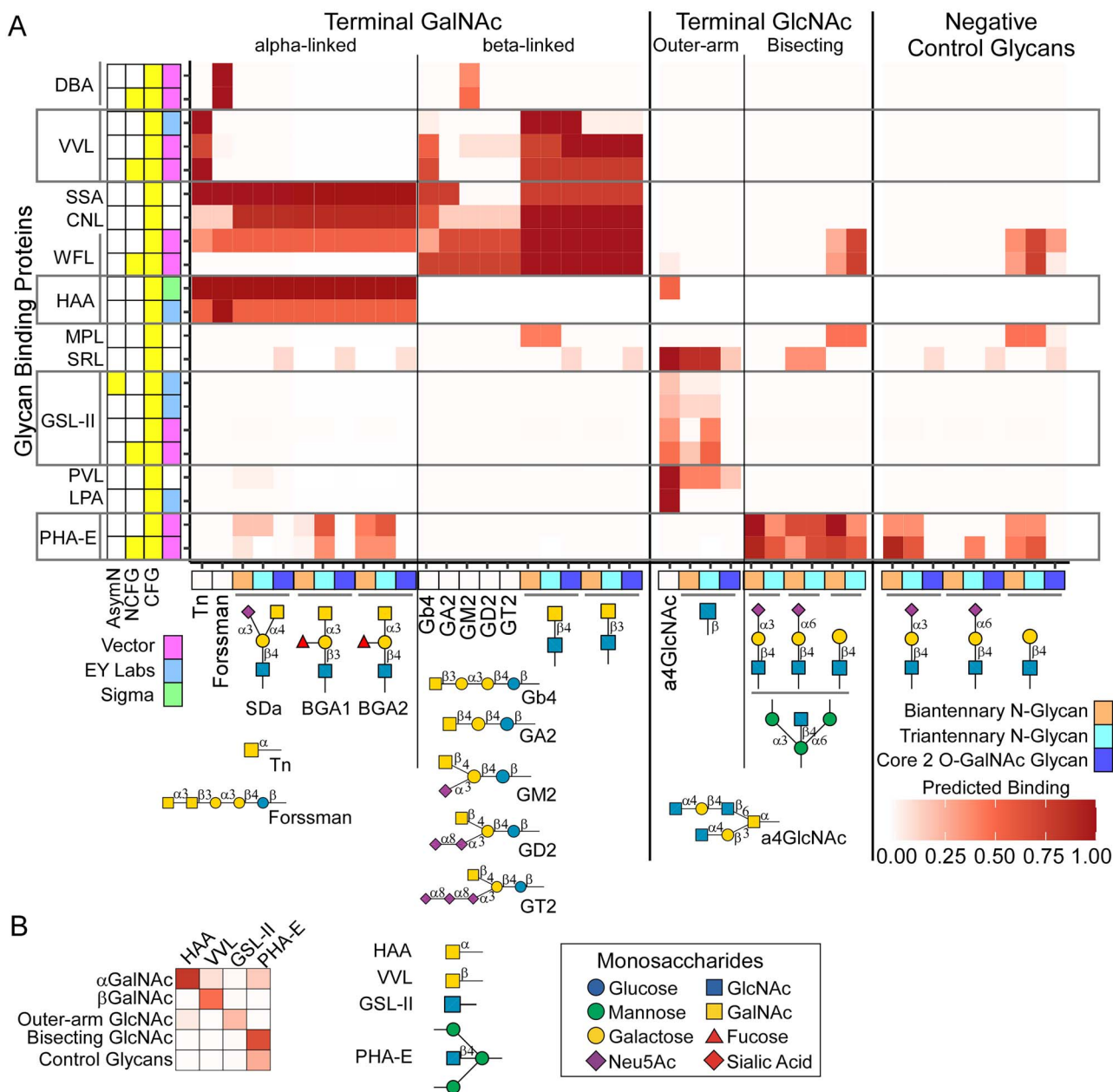


Fig. 6. Searches for glycan-binding proteins that are specific for selected motifs. A) Using models downloaded from CarboGrove that bind any of the comparison motifs, we predicted binding to a series of glycans generated in MotifFinder that contain the motifs, as well as negative-control glycans. B) In the reduced set of 4 lectins (GSL-II, HAA, VVL, and PHA-E), the average binding to the glycans in each group shows a different pattern for each group. The differences in top motifs for the lectins correspond to the differences in binding patterns. Monosaccharide symbols follow the SNFG (Symbol Nomenclature for Glycans) system (PMID 26543186, *Glycobiology* 25: 1323–1324, 2015) details at NCBI.

handle the additional types of information as glycan arrays continue to expand into new areas.

Methods

Data collection

The collaborators at Z Biotech provided 151 glycan-array datasets generated for internal quality control studies, including data for 43 different glycan-binding proteins collected on 11 of the arrays offered by the company. Raw data are available as supplementary data (Supplementary Table III). Details on the data collection are provided in the supplementary MIRAGE document (Supplementary Table II).

Data for the CFG and NCFG arrays were retrieved from their respective websites and databases. Data from individual laboratories were retrieved from the original publications or provided by the authors upon request.

Statistical analysis

The prediction of binding to glycans using the models was described previously (Klamer and Haab 2021) and is detailed in the user’s manual provided with MotifFinder. The reliability score, which was used to rank models by the quality of their results (Fig. 4), measures the difference between the average binding of the top motif and the average binding of the non-binding motifs, normalized by the standard deviation in

the non-binding motifs. This metric is similar to the signal-to-noise ratio. Given the average of the top-binding glycan values m , the average of the non-binding glycan values v , the standard deviation of the non-binding glycan values s , and the number of datasets in the model n , the reliability score is calculated as:

$$\text{Reliability score} = \frac{\sum_1^n (\max(m_n) - v_n) / s_n}{n}$$

Data analysis

The majority of datasets could be analyzed as they were provided, with minor corrections in text syntax to match the modified IUPAC syntax used by the CFG. In one case, the CUPRA array, the data needed to be inverted to enforce the requirement of a positive association between binding and the quantification of the binding. All data were analyzed using MotifFinder release version V2.2.5 (Klamer and Haab 2020).

The Motif Score was calculated as described previously (Klamer et al. 2017). Briefly, a t -test with unequal variances is performed comparing the predicted binding of glycans with the motif to those without the motif. The resulting P -value is log10 transformed and re-signed to match the sign of the t value from the t -test. Motif Score values are truncated to the range of -10 to 10 . The Motif Score is used to rank associations with binding rather than for statistics. Additional metrics are used to break ties in the Motif Score for motif sorting, including the average binding of glycans with the motif and the precision of the motif (the number of glycans with the motif that has a positive predicted binding divided by the number of glycans with the motif).

Assigning standardized IDs to database contents

The assignment of UniProt IDs and PFAM families was done manually through searching the UniProt database (accessed 2022 July 2) (Consortium et al. 2020). Assignment of GlyTouCan IDs to glycans (accessible via the API as part of the `get_data` and `get_result` functions) required several steps. MotifFinder's built-in parser is capable of parsing glycans in their raw format (CFG-like IUPAC condensed). We adapted MotifFinder's glycan printing function to print the standardized IUPAC condensed format for all parsable glycans. The IUPAC condensed format was passed to the GlycanFormat-Converter tool to convert the glycans to WURCS glycan format (Tsuchiya et al. 2019). The GlyCosmos portal offers GlyTouCan search by text to get GlyTouCan IDs (Yamada et al. 2020).

Database Design and Programmatic Access

The database was developed using MariaDB (an open-source MySQL branch) and delivered using php (version 7.1.28) for server-side processing of searches and database interface. Javascript was used for the web interface and the compression of webserver file uploads. Bash scripts and the jq json parser tool were used to process, manage, and return webserver requests. The database and webserver are hosted using Amazon Web Services. An overview of the data ingestion process and API data delivery is given in Supplementary Fig. 2. Briefly, information about the glycan-binding protein, the experimental system, and the data are parsed and entered into multiple

tables. The raw data and the results from the MotifFinder analyses are linked to this information.

Programmatic access to the CarboGrove database is served through an API that allows bioinformatics users to access the database in computational formats and utilize the curated list of glycan binding protein aliases to standardize glycan binding protein names. Details on the use of the API are given in the API page in the CarboGrove website, and a high-level overview of the data-retrieval process is given in Supplementary Fig. 2.

Supplementary material

Supplementary material is available at *GLYCOB Journal* online.

Author contributions

Conceptualization: ZLK, BBH.

Data curation and software: ZLK.

Investigation: CMH, JMB, JEK.

Visualization: ZLK, BBH.

Supervision: BBH, JZ.

Writing—original draft: BBH, ZLK.

Writing—review and editing: BBH, ZLK.

Acknowledgments

We thank Dr Geert-Jan Boons (Heparan sulfate Array), Drs Richard Cummings and Akul Mehta (CFG Array), Dr Ratmir Derda (LiGA), Dr Fabian Pfrengle (Plant Cell Wall Array), and Dr Xuezheng Song (NGGM/NGGM-TM) for sharing data in tabular form and/or providing additional experimental information or advice as it relates to their specific arrays.

Funding

This work was supported by National Institute of General Medical Sciences (R44GM131430, R42GM112750); National Cancer Institute (Early Detection Research Network, [U01CA152653]; Alliance of Glycobiologists for Cancer Research, [U01CA226158]).

Conflict of interest statement

JZ is founder and CEO of Z Biotech. The remaining authors declare that they have no competing interests.

Data availability

The models reported here are available through the CarboGrove website. MotifFinder is available as a standalone tool or through a webserver. The webserver features a simplified interface while running the same algorithm as the standalone program. CarboGrove is licensed under the CC BY-SA 4.0 license and can be accessed at <https://carbogrove.org/>. Raw glycan array data for published datasets are available from their respective publications, and processed data are available through the API or via download in bulk from FigShare: https://figshare.com/articles/dataset/CarboGrove_Glycan_Binding_Data/19274777. Details on the use of CarboGrove API are available under the API homepage: <https://carbogrove.org/api/home.php>. Details on the use of CarboGrove and use cases for CarboGrove and its API are available through the help page: <https://carbogrove.org/help.php>. Unpublished glycan-array data are available in the supplementary materials. The compiled MotifFinder software used to generate all results in CarboGrove is freely available

for academic, nonprofit, and research use through our download page (<https://haablab.vai.org/tools/>).

Abbreviations

AAL, *Aleuria Aurantia* Lectin; API, Application Programming Interface; CCL2, *Coprinopsis Cinerea* Lectin 2; CFG, Consortium for Functional Glycomics; CEMA, Chemoenzymatic Modular Assembly; GalNAc, N-acetyl galactosamine; GlcA, glucuronic acid; GlcNAc, N-acetyl glucosamine; GlyMDB, Glycan Microarray Database; GLAD, Glycan Array Dashboard; GSL-II, *Griffonia simplicifolia* Lectin 2; HAA, *Helix Aspersa* Agglutinin; LacDiNAc, GalNAc, GlcNAc disaccharide; LacNAc, Galactose, GlcNAc disaccharide; LiGA, Liquid Glycan Array; LCA, *Lens Culinaris* Agglutinin; LFDB, Lectin Frontier Database; LTA, *Lotus Tetragonolobus* Lectin; MCAW-DB, Multiple Carbohydrate Alignment with Weights Database; PHA-E, *Phaseolus Vulgaris* Erythroagglutinin; Neu5Ac, 5-acetyl neuramic acid; Neu5Gc, 5-glycolyl neuramic acid; NCFG, National Center for Functional Glycomics; NGGM, Next Generation Glycan Microarray; NGGM-TM, Next Generation Glycan Microarray Traditional Microarray Comparison; SNA, *Sambucus Nigra* Agglutinin 1; WGA, Wheat Germ Agglutinin

References

- Blixt O, Head S, Mondala T, Scanlan C, Huflejt ME, Alvarez R, Bryan MC, Fazio F, Calarese D, Stevens J, *et al.* Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc Natl Acad Sci U S A*. 2004;101(49):17033–17038.
- Bojar D, Meche L, Meng G, Eng W, Smith DF, Cummings RD, Mahal LK. A useful guide to lectin binding: machine-learning directed annotation of 57 unique lectin specificities. *ACS Chem Biol*. 2022. doi: <https://doi.org/10.1021/acscchembio.1c00689>.
- Bonnardel F, Mariethoz J, Salentin S, Robin X, Schroeder M, Perez S, Lisacek F, Imberty A. UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. *Nucleic Acids Res*. 2019;47(D1):D1236–D1244.
- Bryan MC, Fazio F, Lee H-K, Huang C-Y, Chang A, Best MD, Calarese DA, Blixt O, Paulson JC, Burton D, *et al.* Covalent display of oligosaccharide arrays in microtiter plates. *J Am Chem Soc*. 2004;126(28):8640–8641.
- Büll C, Nason R, Sun L, Coillie JV, Sørensen DM, Moons SJ, Yang Z, Arbitman S, Fernandes SM, Furukawa S, *et al.* Probing the binding specificities of human Siglecs by cell-based glycan arrays. *Proc National Acad Sci*. 2021;118(17):e2026102118. doi: <https://doi.org/10.1073/pnas.2026102118>.
- Cao Y, Park S-J, Mehta AY, Cummings RD, Im W. GlyMDB: glycan microarray database and analysis toolset. *Bioinformatics*. 2019;36(8): 2438–2442.
- Cholleti SR, Agravat S, Morris T, Saltz JH, Song X, Cummings RD, Smith DF. Automated motif discovery from glycan array data. *Omics J Integr Biology*. 2012;16(10):497–512.
- Chopra P, Joshi A, Wu J, Lu W, Yadavalli T, Wolfert MA, Shukla D, Zaia J, Boons G-J. The 3-O-sulfation of heparan sulfate modulates protein binding and lyase degradation. *Proc National Acad Sci*. 2021;118(3):e2012935118. doi: <https://doi.org/10.1073/pnas.2012935118>.
- Coff L, Chan J, Ramsland PA, Guy AJ. Identifying glycan motifs using a novel subtree mining approach. *BMC Bioinformatics*. 2020;21(1):42. doi: <https://doi.org/10.1186/s12859-020-3374-4>.
- Consortium TU, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2020;49(D1):D480–D489.
- Fujita A, Aoki NP, Shinmachi D, Matsubara M, Tsuchiya S, Shiota M, Ono T, Yamada I, Aoki-Kinoshita KF. The international glycan repository GlyTouCan version 3.0. *Nucleic Acids Res*. 2020;49(D1):D1529–D1533.
- Geissner A, Reinhardt A, Rademacher C, Johannsen T, Monteiro J, Lepenies B, Thépaut M, Fieschi F, Mrázková J, Wimmerova M, *et al.* Microbe-focused glycan array screening platform. *Proc Natl Acad Sci*. 2019;116(6):201800853. doi: <https://doi.org/10.1073/pnas.1800853116>.
- Grant OC, Smith HM, Firsova D, Fadda E, Woods RJ. Presentation, presentation! Molecular-level insight into linker effects on glycan array screening data. *Glycobiology*. 2014;24(1):17–25.
- Hinou H, Kikuchi S, Ochi R, Igarashi K, Takada W, Nishimura S-I. Synthetic glycopeptides reveal specific binding pattern and conformational change at O-mannosylated position of α -dystroglycan by POMGnT1 catalyzed GlcNAc modification. *Bioorg Med Chem*. 2019;27(13):2822–2831.
- Hirabayashi J, Tateno H, Shikanai T, Aoki-Kinoshita KF, Narimatsu H. The Lectin Frontier Database (LfDB), and data generation based on frontal affinity chromatography. *Molecules*. 2015;20(1):951–973.
- Horton M, Su G, Yi L, Wang Z, Xu Y, Pagadala V, Zhang F, Zaharoff DA, Pearce K, Linhardt RJ, *et al.* Construction of heparan sulfate microarray for investigating the binding of specific saccharide sequences to proteins. *Glycobiology*. 2020;31(3):188–199.
- Hosoda M, Takahashi Y, Shiota M, Shinmachi D, Inomoto R, Higashimoto S, Aoki-Kinoshita KF. MCAW-DB: a glycan profile database capturing the ambiguity of glycan recognition patterns. *Carbohydr Res*. 2018;464:44–56.
- Kitov PI, Kitova EN, Han L, Li Z, Jung J, Rodrigues E, Hunter CD, Cairo CW, Macauley MS, Klassen JS. A quantitative, high-throughput method identifies protein–glycan interactions via mass spectrometry. *Commun Biol*. 2019;2(1):268. doi: <https://doi.org/10.1038/s42003-019-0507-2>.
- Kitova EN, Han L, Vinals DF, Kitov PI, Derda R, Klassen JS. Influence of labeling on the glycan affinities and specificities of glycan-binding proteins. A case study involving a C-terminal fragment of human galectin-3. *Glycobiology*. 2019;30(1):49–57.
- Klamer Z, Haab B. 2020. Automated identification of lectin fine specificities from glycan-array data. Vol. 1346. American Chemical Society. (*Glycan-Based Cellular Communication: Techniques for Carbohydrate-Protein Interactions*). p. 67–82. doi: <https://doi.org/10.1021/bk-2019-1346.ch005>.
- Klamer Z, Haab B. Combined analysis of multiple glycan-array datasets: new explorations of protein–glycan interactions. *Anal Chem*. 2021;93(31):10925–10933.
- Klamer Z, Staal B, Prudden AR, Liu L, Smith DF, Boons G-J, Haab B. Mining high-complexity motifs in glycans: a new language to uncover the fine specificities of lectins and glycosidases. *Anal Chem*. 2017;89(22):12342–12350.
- Li L, Guan W, Zhang G, Wu Z, Yu H, Chen X, Wang PG. Microarray analyses of closely related glycoforms reveal different accessibilities of glycan determinants on N-glycan branches. *Glycobiology*. 2019a;30(5):334–345.
- Li T, Liu L, Wei N, Yang J-Y, Chapla DG, Moremen KW, Boons G-J. An automated platform for the enzyme-mediated assembly of complex oligosaccharides. *Nat Chem*. 2019b;11(3):229–236.
- Li C, Palma AS, Zhang P, Zhang Y, Gao C, Silva LM, Li Z, Trovão F, Weishaupt M, Seeberger PH, *et al.* Non-covalent microarrays from synthetic amino-terminating glycans-implications in expanding glycan microarray diversity and platform comparison. *Glycobiology*. 2021;31(8):931–946.
- Liu Y, McBride R, Stoll M, Palma AS, Silva L, Agravat S, Aoki-Kinoshita KF, Campbell MP, Costello CE, Dell A, *et al.* The minimum information required for a glycomics experiment (MIRAGE) project: improving the standards for reporting glycan microarray-based data. *Glycobiology*. 2017;27(4):280–284.
- Mehta AY, Heimbürg-Molinario J, Cummings RD. Tools for generating and analyzing glycan microarray data. *Beilstein J Org Chem*. 2020;16(1):2260–2271.
- Mende M, Tsouka A, Heidepriem J, Paris G, Mattes DS, Eickelmann S, Bordoni V, Wawrzinek R, Fuchsberger FF, Seeberger PH, *et al.* On-chip neo-glycopeptide synthesis for multivalent glycan presentation. *Chem Eur J*. 2020;26(44):9954–9963.

- Nason R, Büll C, Konstantinidi A, Sun L, Ye Z, Halim A, Du W, Sørensen DM, Durbeson F, Furukawa S, *et al.* Display of the human mucinome with defined O-glycans by gene engineered cells. *Nat Commun.* 2021;12(1):4070. doi: <https://doi.org/10.1038/s41467-021-24366-4>.
- Padler-Karavani V, Song X, Yu H, Hurtado-Ziola N, Huang S, Muthana S, Chokhawala HA, Cheng J, Verhagen A, Langereis MA, *et al.* Cross-comparison of protein recognition of sialic acid diversity on two novel sialoglycan microarrays. *J Biol Chem.* 2012;287(27):22593–22608.
- Porter A, Yue T, Heeringa L, Day S, Suh E, Haab BB. A motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins. *Glycobiology.* 2010;20(3):369–380.
- Prudden AR, Liu L, Capicciotti CJ, Wolfert MA, Wang S, Gao Z, Meng L, Moremen KW, Boons G-J. Synthesis of asymmetrical multiantennary human milk oligosaccharides. *Proc Natl Acad Sci.* 2017;114(27):6954–6959.
- Ruprecht C, Bartetzko MP, Senf D, Dallabernadina P, Boos I, Andersen MCF, Kotake T, Knox JP, Hahn MG, Clausen MH, *et al.* A synthetic glycan microarray enables epitope mapping of plant cell wall glycan-directed antibodies. *Plant Physiol.* 2017;175(3):1094–1104.
- Sojitra M, Sarkar S, Maghera J, Rodrigues E, Carpenter EJ, Seth S, Vinals DF, Bennett NJ, Reddy R, Khalil A, *et al.* Genetically encoded multivalent liquid glycan array displayed on M13 bacteriophage. *Nat Chem Biol.* 2021;17(7):806–816.
- Song X, Yu H, Chen X, Lasanajak Y, Tappert MM, Air GM, Tiwari VK, Cao H, Chokhawala HA, Zheng H, *et al.* A sialylated glycan microarray reveals novel interactions of modified sialic acids with proteins and viruses. *J Biol Chem.* 2011;286(36):31610–31622.
- Temme JS, Campbell CT, Gildersleeve JC. Factors contributing to variability of glycan microarray binding profiles. *Faraday Discuss.* 2019;219:90–111.
- Tsuchiya S, Yamada I, Aoki-Kinoshita KF. GlycanFormatConverter: a conversion tool for translating the complexities of glycans. *Bioinformatics.* 2019;35(14):2434–2440.
- Wang L, Cummings RD, Smith DF, Huflejt M, Campbell CT, Gildersleeve JC, Gerlach JQ, Kilcoyne M, Joshi L, Serna S, *et al.* Cross-platform comparison of glycan microarray formats. *Glycobiology.* 2014;24(6):507–517.
- Wang S, Chen C, Guan M, Liu D, Wan X-F, Li L. terminal epitope-dependent branch preference of Siglecs toward N-glycans. *Front Mol Biosci.* 2021;8:645999. doi: <https://doi.org/10.3389/fmolb.2021.645999>.
- Yamada I, Shiota M, Shinmachi D, Ono T, Tsuchiya S, Hosoda M, Fujita A, Aoki NP, Watanabe Y, Fujita N, *et al.* The GlyCosmos portal: a unified and comprehensive web resource for the glycosciences. *Nat Methods.* 2020;17(7):649–650.
- York WS, Mazumder R, Ranzinger R, Edwards N, Kahsay R, Aoki-Kinoshita KF, Campbell MP, Cummings RD, Feizi T, Martin M, *et al.* GlyGen: computational and informatics resources for glycoscience. *Glycobiology.* 2019;30(2):72–73.
- Zhang J, Chen C, Gadi MR, Gibbons C, Guo Y, Cao X, Edmunds G, Wang S, Liu D, Yu J, *et al.* Machine-driven enzymatic oligosaccharide synthesis by using a peptide synthesizer. *Angew Chem Int Ed.* 2018;57(51):16638–16642.