

Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Sanjay B. Agravat

Date _____

Bioinformatics Methods and Tools for Glycomics

By

Sanjay B. Agravat
Doctor of Philosophy

Computer Science and Informatics

Lee Cooper, Ph.D.
Advisor

Richard Cummings, Ph.D.
Advisor

David Gutman, M.D., Ph.D.
Committee Member

David Smith, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date _____

Bioinformatics Methods and Tools for Glycomics

by

Sanjay B. Agravat

B.S., Florida State University, 1999

M.S., Johns Hopkins University, 2007

Advisor: Lee Cooper, Ph.D.

Advisor: Richard Cummings, Ph.D.

An abstract of

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Computer Science and Informatics

2015

Abstract

Bioinformatics Methods and Tools for Glycomics

By Sanjay B. Agravat

Glycomics is the study of the structure and function of carbohydrates in biological systems. In comparison to the expansion of the more established fields of genomics and proteomics, the integration of glycans and glycomics in biomedical research has lagged far behind. Glycomics has the potential to be included as another foundational science in the study of human disease, since glycans play major roles in certain hereditary diseases, infectious diseases, and cancer. The structural and functional complexity of glycans coupled with the lack of robust bioinformatics impedes the integration of glycoscience into the scientific mainstream. The central objective of this thesis is to develop novel computational methods and bioinformatics tools to advance the understanding of structure and function relationships of glycans and their recognition and binding by Glycan Binding Proteins (GBPs).

We have developed a method to automate the interpretation of glycan microarray data to identify the glycan determinants that are necessary for binding. We evaluate this method against GBPs of known specificities to validate the results. We demonstrate this approach revealed new recognition motifs that had not been previously reported.

We also present a novel computational approach to automate the sequencing of glycans based on a method known as “Metadata-Assisted Glycan Sequencing” (MAGS), which combines analyses of glycan structures by mass spectrometry (MS) and glycan microarray technology to fully characterize glycan sequences. We target the soluble glycans in the human milk glycome as the first meta-glycome to be defined using this method.

To facilitate access by scientists to glycomics information, we developed an open-source web-based bioinformatics platform for glycan microarray analysis. The platform provides interactive visualization features to view, search, and compare experimental data and also includes glycan motif mining and analysis.

In addressing these research areas, we have developed novel methods, algorithms, and software tools applied to the field of glycomics. These contributions will aid in the elucidation of the human glycome and a greater understanding of the diverse and important biological functions of glycans.

Bioinformatics Methods and Tools for Glycomics

by

Sanjay B. Agravat

B.S., Florida State University, 1999

M.S., Johns Hopkins University, 2007

Advisor: Lee Cooper, Ph.D.

Advisor: Richard Cummings, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Computer Science and Informatics

2015

Acknowledgements

I would like to express my appreciation and gratitude to many people that have supported me throughout the journey of my PhD. I would like to thank Prof. Richard Cummings for his mentoring and support of my PhD research. I am grateful to have received my scientific training from you and I appreciate the personal and professional encouragement you gave me throughout these years. You have my deep respect for your contributions to science which paved the way for my research. My sincere thanks to Dr. Lee Cooper for his patience and his countless support and guidance throughout my journey as a PhD student. You were always willing to provide me with encouragement and mentoring even before I became your advisee.

Special thanks to Dr. David Smith whom I have collaborated with for several years. I learned a great deal about biochemistry from you and I am also grateful for the scientific mentoring you gave me. Your advocacy for bioinformatics in glycomics has opened big doors for me and I am very thankful for your support and vision. I would also like to thank the NIH funding agency for their support of the National Center for Functional Glycomics which I worked at during my PhD.

Many thanks to Dr. David Gutman for agreeing to serve on my dissertation committee and his review of my work. There are many faculty and former faculty members from Emory University I have worked with that have influenced me in my research including Dr. Patrick Widener and Dr. Tahsin Kurc in the area of high performance computing, Dr. Carlos Moreno for his work in cancer bioinformatics, Dr. James Nettles for his work in computational chemistry, and Dr. James Taylor for his work in bioinformatics and reproducible research. Special thanks to Dr. Joel Saltz for advising me in my early years as a PhD student. I gained a broad and deep exposure to machine learning, high performance computing, and biomedical informatics from working in his lab. My sincere thanks to the members of the Cummings Laboratory, including Hong Ju, Yi Lasanajak, Xuezheng Song, and Ying Yu for the experimental analysis they conducted which ultimately produced the data I analyzed.

Special thanks to Dr. Ashish Sharma and Dr. Andrew Post for their encouragement and supporting me with work opportunities during my PhD. Many thanks to the other faculty in the BMI department including Dr. Jun Kong and Dr. Shamim Nemati for making the BMI department a congenial atmosphere to work in.

To my parents, Dr. and Mrs. Bansidas and Gauri Agravat, you instilled in me respect, discipline, and hard work. You provided me with everything I needed to be successful even though you both came from humble beginnings that I could not

fathom. My brothers Manoj and Jaydeep, I also thank you for your support and encouragement over the years.

My children Suhina and Sujaan, you are my inspiration and what I live for. You put a smile on my face everyday. I hope you both find happiness and something in life you are passionate about and work hard to achieve your goals.

Finally, to my dear wife, Namita, you persevered through this PhD just like I did. You sacrificed and never wavered in your support for me to pursue my academic interests. There are no words to express how grateful I am for your support so I will just have to take a lifetime to thank you.

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Background on Glycobiology and Glycome Informatics	3
1.2.1	Glycan Structures	3
1.2.2	Glycan Classifications	5
1.2.3	Glycan Nomenclature	6
1.2.4	Protein-Glycan Interactions	9
1.2.5	Bioinformatics for Glycan Microarray Data	12
1.2.6	Mining for Glycan Motifs on Glycan Microarray Data	14
1.2.7	Deciphering and Sequencing the Human Glycome	16
1.3	Contributions of This Thesis	17
2	Automated Motif Discovery from Glycan Microarray Data	19
2.1	Introduction	19
2.2	Materials and Methods	21
2.3	Results	27
2.4	Discussion	45
3	Computational Approaches to Define the Human Milk Metaglycome	54
3.1	Introduction	54
3.2	Methods	57

3.3	Results	61
3.4	Discussion	67
3.5	Conclusion	71
4	A Web-Based Bioinformatics Platform for Glycan Microarray Analysis	73
4.1	Introduction	73
4.2	Methods	75
4.3	Results	78
4.4	Discussion	82
5	Conclusion and Future Work	83
5.1	Future Work	83
	Bibliography	85

Figures

2.1	Subtree mining	26
2.2	Summary of GLYMMR	28
2.3	Display of motifs for SNA	29
2.4	SNA weak binding	34
2.5	Discriminatory capability of GLYMMR	35
2.6	Display of motifs for HPA	37
2.7	Display of motifs for Con A	39
2.8	Display of motifs for Con A for two concentrations	41
2.9	Display of motifs for UEA-I	43
2.10	Display of motifs for Gal-8	46
3.1	GlycomeSeq Algorithm	62
3.2	Virtual Glycome	64
3.3	Terminal Determinants	66
3.4	GlycomeSeq display for HMG-20	68
4.1	Experiment chart	79
4.2	Heatmap for AAL and UEA-I	80
4.3	Glycan search	81

Chapter 1

Introduction

1.1 Introduction

Carbohydrates are one of the four fundamental classes of biological macromolecules, along with nucleic acids, proteins, and lipids. Carbohydrates, also known as glycans, play an integral role in essential biological functions including cell signaling, molecular recognition, immunity, and inflammation. All living cells are coated with glycans on their cell surface and these glycans serve as interaction targets for pathogens and other biochemical processes. Glycans play a critical role across all aspects of human health and disease [106]. More than half of all human proteins are post-translationally modified with glycans, a process known as glycosylation. Proteins essential to normal cell physiology are usually glycosylated and variation in their glycosylation patterns often leads to changes in their function and contributes to disease [8].

In the post-genomic era, a complex systems level approach has emerged to understanding and treating human disease [61]. Yet, mention of glycomics alongside transcriptomics, proteomics, metabolomics, and other ‘omics in this systems approach is typically uncommon. Compared to the molecular biology revolution in the 1960s and 1970s, the study of glycomics lagged far behind due to the structural complexity and complicated method of biosynthesis relative to genes and proteins [109]. The challenges to the structure-function understanding of glycans is summarized in fig-

ure 1.1.

The complexity of glycan structures and glycan synthesis, relative to nucleic acids and proteins, presents a challenge in experimental and computational analysis of glycan structure and function relationships. Oligosaccharide chains often form branched structures and each monosaccharide residue can be coupled to different positions on adjacent monosaccharides in α or β anomeric configurations. Moreover, individual sugar moieties within a chain can be modified by phosphorylation, sulfation, methylation, O-acetylation, or fatty acylation, further increasing the combinatorial possibilities for oligosaccharide diversity. Partially owing to this high degree of complexity, the field of Glycoscience has been relatively understudied.

Meanwhile, the development of databases and software tools accelerated the efforts of the human genome-sequencing project:

“The wide availability of genome sequence data has created a wealth of opportunities, most notably in the realm of functional genomics and proteomics. This quiet revolution in the biological sciences has been enabled by our ability to collect, manage, analyze, and integrate large quantities of data. In the process, bioinformatics has itself developed from something considered to be little more than information management and the creation of sequence-search tools into a vibrant field encompassing both highly sophisticated database development and active pure and applied research programs in areas far beyond the search for sequence homology.”[79]

In order to broaden access of glycomics to the entire scientific community, there must be development of bioinformatics resources and tools to make glycomics information and data more accessible. What is known about the function of glycans and the details of the human glycome is still incomplete but with the advances in chemical and chemoenzymatic synthesis of glycans [16] and tools that have facilitated decoding the structural heterogeneity of the glycome [91], we can begin to decipher the human glycome. As recent research indicates “glycans are directly involved in the pathophysiology of every major disease” and that “additional knowledge from glycoscience

will be needed to realize the goals of personalized medicine and to take advantage of the substantial investments in human genome and proteome research and its impact on human health”, therefore it is critical to integrate glycomics as another toolkit for ‘omics [35].

This research is focused on computational methods and bioinformatics tools to advance the understanding of structure and function relationships of glycans and their recognition and binding by Glycan Binding Proteins (GBPs). The rest of the chapter is organized by background information on glycan structures, glycan terminology, glycan-protein interactions, and a review of the existing methods and tools relevant to this research topic.

1.2 Background on Glycobiology and Glycome Informatics

The complexity of glycan structures presents a challenge in the representation, storage, and extraction of information from glycomics experimental data. This section will provide an overview of glycan structures and a background in glycome informatics relevant to the representation, storage, and analysis of glycan structures. The section on *Glycan Structures* and *Glycan Classifications* appears in Chapter 1 of “The Essentials of Glycobiology” [106].

1.2.1 Glycan Structures

Monosaccharides are the fundamental units of carbohydrates that can form covalent bonds to other monosaccharides. Due to chemical stability, the most common classification of monosaccharides are the five carbon membered pentose (or furanose) rings and the six carbon membered hexose (or pyranose) rings. There are approximately 20 monosaccharides found in the mammalian glycome. The ringed form of the monosaccharide leads to an asymmetric center known as the anomeric carbon. There are two further stereoisomers that may be formed due to the orientation of

Key challenge	Features	Impact on study of glycans
Glycan biosynthesis	Nontemplate-driven process, unlike DNA/RNA and protein	Replication- or translation-like "rules" cannot be easily applied; no direct methods to amplify glycans, unlike DNA (PCR) and protein (recombinant expression)
	Limited availability of glycans from natural sources (e.g., cells, tissues)	Without amplification tools, analytical and functional methods often require high sensitivity
	Tissue-, development-, and metabolic-dependent expression of glycan biosynthetic machinery (glycosyltransferases and glycosidases)	Glycan structure is sensitive to cellular conditions, tissue type, and developmental stages
	Lack of proofreading in glycan biosynthetic process	Increases structural diversity of glycans to be analyzed
Glycan structural complexity and heterogeneity	Presence of isomers and different anomeric configurations	Properties generally not present in DNA and proteins; challenges structural characterization by single method
	Microheterogeneity—a range of glycan structures (length, composition, branching) found at any given glycosylation site on a glycoprotein	Highly similar physicochemical properties of glycan microheterogeneities challenges their characterization
	Branching	Unambiguous designation of branches and their locations challenged by analytical approaches
	Presence of multiple modifications (sulfation, acetylation, methylation) and high diversity of linkages (location of linkages and anomericity)	Chemical synthesis is difficult and limited to small oligosaccharides due to the need of complex protecting and deprotecting strategies
	Site of attachment to protein/lipid	Requires glycan–protein and/or glycan–lipid characterization in addition to glycan structure
Glycan presentation and interactions	Presentation of an ensemble of different (often related) structures within a biological system or interaction	Studies must account for a population of glycans with similar structures, rather than an "average" single structure
	Glycan–protein interactions often achieve high affinity and specificity by multivalency	Correct presentation of glycan and glycan-binding protein/domain(s) is critical for experimental design
	Glycan–protein interactions modulate biology in an analog-like nature	Functional readouts must be characterized in terms of gradation of effects (not binary "on/off" effects)
	High torsional flexibility of glycans mediates presentation of a range of conformations for a particular glycan	Sequence of glycan is often not sufficient to characterize glycan–protein interactions; analysis of conformations and topologies should be considered

Figure 1.1: Challenges to the structure-function understanding of glycans [84]

the anomeric carbon and the stereogenic center farthest away from it. The alpha anomer is defined as having the same configuration between the anomeric carbon and the stereogenic center, while the beta anomer is defined as having a different configuration.

Two monosaccharides may have the same mass, thus same chemical formula, and still be different structures, known as isomers. These isomers differ in their orientation and are considered stereoisomers. Three common hexose isomers found in mammals are galactose, glucose, and mannose. Each isomer can further be classified by its chirality, either the D or L configuration (i.e. D-Glucose L-Glucose); mammalian monosaccharides commonly exist in the D configuration.

Monosaccharides may be connected by glycosidic bonds to form disaccharides and polysaccharides. These linkages are formed by releasing a molecule of water between sugars to form a glycoside. Unlike DNA and amino acid sequences, linkages from monosaccharides may be connected to more than one other monosaccharide unit, to form a branched structure.

Carbohydrates can be attached to proteins to form glycoproteins. The carbohydrates of glycoproteins are commonly referred to as glycans, which are assembled from simple sugars. A glycosidic bond links a carbohydrate to the side chain of asparagine (N-linked) or to the side chain of serine or threonine (O-linked).

1.2.2 Glycan Classifications

Carbohydrates can be attached to proteins to form glycoproteins. The carbohydrates of glycoproteins are commonly referred to as glycans, which are assembled from simple sugars. Mammalian glycans can be characterized by four broad classifications determined by the core structure of the glycan at the reducing end of the glycoprotein.

N-Glycans. The most common class of glycans found in mammals is the N-Glycan. A glycosidic bond from the reducing end of the core structure linked to the side chain of asparagine is referred as an N-Linked Glycan. These glycans share a common pentasaccharide core for three major subclasses: high-mannose, complex, and hybrid. N-Glycans are important for protein folding and play a role in cell interactions.

O-Glycans. An O-glycan is typically linked to the glycoprotein via N-acetylgalactosamine (GalNAc) to a serine or threonine residue and can be extended into a variety of different structures. O-glycans are relatively smaller structures and can be found on large glycoproteins called mucin that binds to pathogens as part of the immune system.

Glycosphingolipids (GSLs). A glycosphingolipid (often called glycolipid) is an oligosaccharide usually attached via glucose or galactose to a ceramide residue, which is itself composed of a long chain base and a fatty acid. There are several types of glycolipids found in animal cells including the lacto series found in human milk.

Glycosoaminoglycans (GAGs). Glycosaminoglycans are long, linear polysaccharides containing a repeating disaccharide unit of two modified sugars, N-acetylgalactosamine (GalNAc) or N-acetylglucosamine (GlcNAc), and a uronic acid such as glucuronate or iduronate. GAGs are highly negatively charged molecules, with extended conformation that imparts high viscosity to the solution. GAGs are located primarily on the surface of cells or in the extracellular matrix (ECM).

1.2.3 Glycan Nomenclature

The necessity to represent glycan structures for scientific communication is challenging due to the variations in structural characteristics and connectivity between monosaccharides. The use of symbols to depict glycans originated from Kornfeld in 1978, was systematized in the First Edition of “Essentials of Glycobiology” and

updated for the second edition, with input from relevant organizations such as the Consortium for Functional Glycomics [107]. The representation of monosaccharides as symbols and textual codes has various different formats due to independent development of glycan databases around the same time [4]. The common set of descriptors for glycan structures include the residue, modifications, substitutions, absolute configuration, anomer configuration, glycosidic linkages, and ring conformation. The inherent branched nature of glycans structures coupled with this set of descriptors, makes for a complex encoding scheme.

Due to the branched nature of glycans, a tree-like or even graph-like format could be used as a representation for a glycan structures; however, many nomenclatures adopted a linearized encoding of glycan structures that use special characters to represent branching. The linear formats have the advantage of a condensed text representation compared to a more verbose representation such as XML. The most common ways of encoding and representing glycans are described in Chapter 2 of “Glycome Informatics: Methods and Applications” [4]. A summary of selected nomenclatures is described below and illustrated in figure 1.2.

1.2.3.1 Extended IUPAC

The extended IUPAC Nomenclature of Carbohydrates is a linear format that uses a three-letter code to represent monosaccharides with each code preceded by the anomeric descriptor and the absolute configuration symbol. The linkages between the carbon atoms from two monosaccharides are indicated by the carbon position of the linkage separated by a dash or arrow and are located between parentheses. Branching is indicated by square brackets.

1.2.3.2 Modified IUPAC condensed

Similar to the extended IUPAC except it includes the anomeric carbon but not the parentheses, and can be written in either a linear or 2D representation. Branching is indicated by parentheses.

1.2.3.3 Linear Code

Single letter nomenclature for monosaccharides and includes the anomeric carbon of the parent to which the residue is attached. Branching is indicated by parentheses. The default format is based on common configurations of glycan structures (i.e. D vs. L configurations) but also has rules to deal with repeating units, cycles, uncommon configurations, and unknown linkages, anomers, and residues.

1.2.3.4 KEGG Chemical Function (KCF)

A graph notation format where nodes are monosaccharides and edges are glycosidic linkages. The format also permits an optional name for the glycan, composition, mass, repeating units, and more. The format requires a label for a node followed by the residue name and the x and y coordinate by which to draw the structure in 2D.

1.2.3.5 GlycoCT

A graph representation that has a condensed format and an XML format. The condensed format allows for unique identification of glycan structures while the XML format facilitates data exchange. The format was designed to handle the full complexity of experimentally derived structural carbohydrate sequence data across all taxonomic sources.

1.2.3.6 GLYDE-II

A graph representation that was developed as a standard for data exchange of carbohydrate structures in an XML format. The format allows the complete structure of biological molecules to be completely and unambiguously specified at several levels of granularity

1.2.3.7 CFG Symbolic Notation

A symbolic notation where isobaric sugars share the same shape and isomeric sugars are differentiated by color. Similar derivatives also share the same color. Colored symbols should still appear distinguishable when copied or printed in black and white. The symbolic key of monosaccharides is illustrated in figure 1.3.

1.2.4 Protein-Glycan Interactions

Glycans are recognized by various glycan-binding proteins (GBPs), which include glycosyltransferases, antibodies, animal and plant lectins, glycosaminoglycan-binding proteins, bacterial adhesins, and viral hemagglutinins. GBPs often have multiple glycosylation sites with high specificity for glycans. While these are typically low affinity interactions, the avidity effects of multivalency allow for stronger binding. Protein-Glycan interactions are also bounded by the same principles as other macromolecular interactions such as hydrogen bonding, hydrophobic interactions, and electrostatic interactions. GBP binding sites may accommodate glycan determinants made up of two to six linear monosaccharides that can also accommodate side-chain modifications.

It is estimated there are over 7,000 glycan determinants in the human glycome [27]. These determinants play important functional roles in the cellular environment including binding and signaling receptors. Sophisticated techniques have been de-

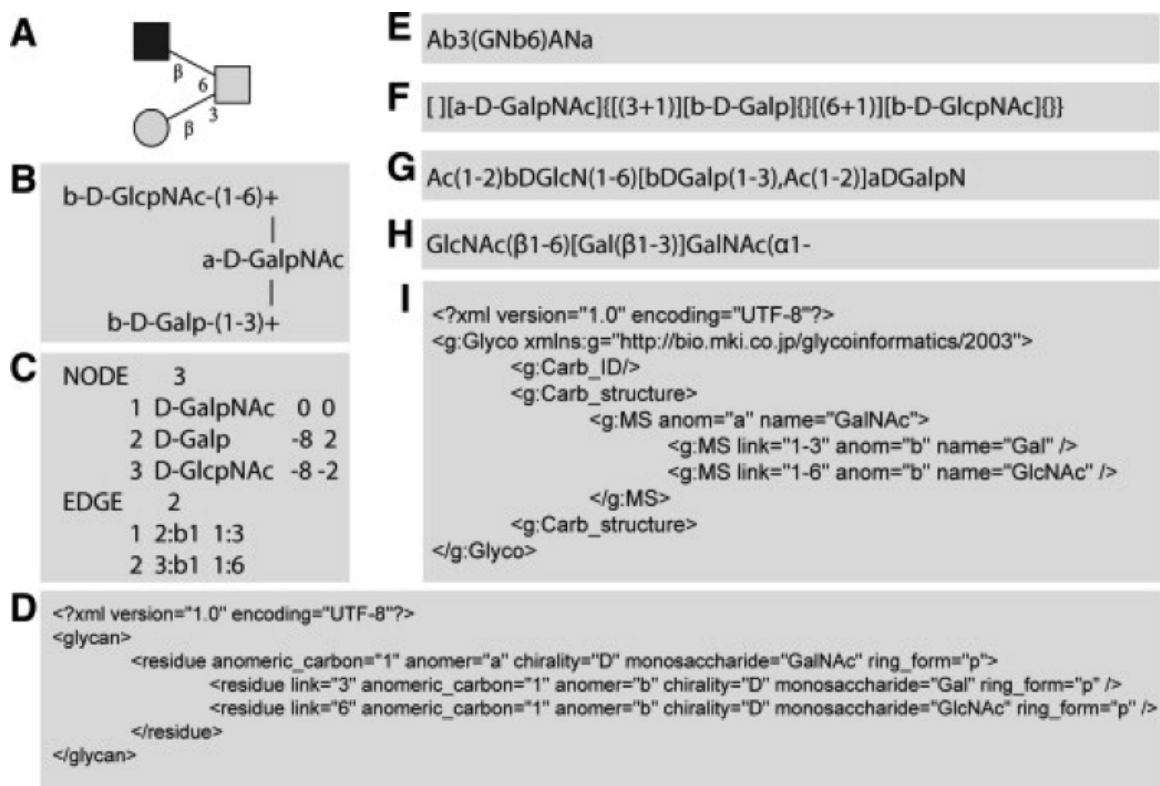


Figure 1.2: An overview of the sequence formats used in glycobioinformatics, using as an example the O-glycan core 2 motif. Normally, database users will be confronted with graphical representations similar to A and B. (A) Graphical representation as suggested by the Consortium for Functional Glycomics (<http://glycomics.scripps.edu/CFGnomenclature.pdf>). (B) Carbohydrate Bank two-dimensional graph. (C) KCF, an application of the connection table approach, as used by the Kyoto Encyclopedia of Genes and Genomes (KEGG). (D) Glycan data exchange format (Glyde), an XML variant. (E) GlycoMinds encoding, as employed by the CFG15 (sequences are read from right to left, AN is d-GalpNAc, GN is d-GlcpNAc, A is d-Galp. An online introduction is available 16). (F) LINUCS sequence format as applied in GLY-COSCIENCES.de. (G) Bacterial carbohydrate structure database (BCSDB) encoding. (H) GlycoSuiteDB format. (I) CabosML, another XML-variant. Reference: <http://dx.doi.org/10.1016/j.carres.2008.03.011>.

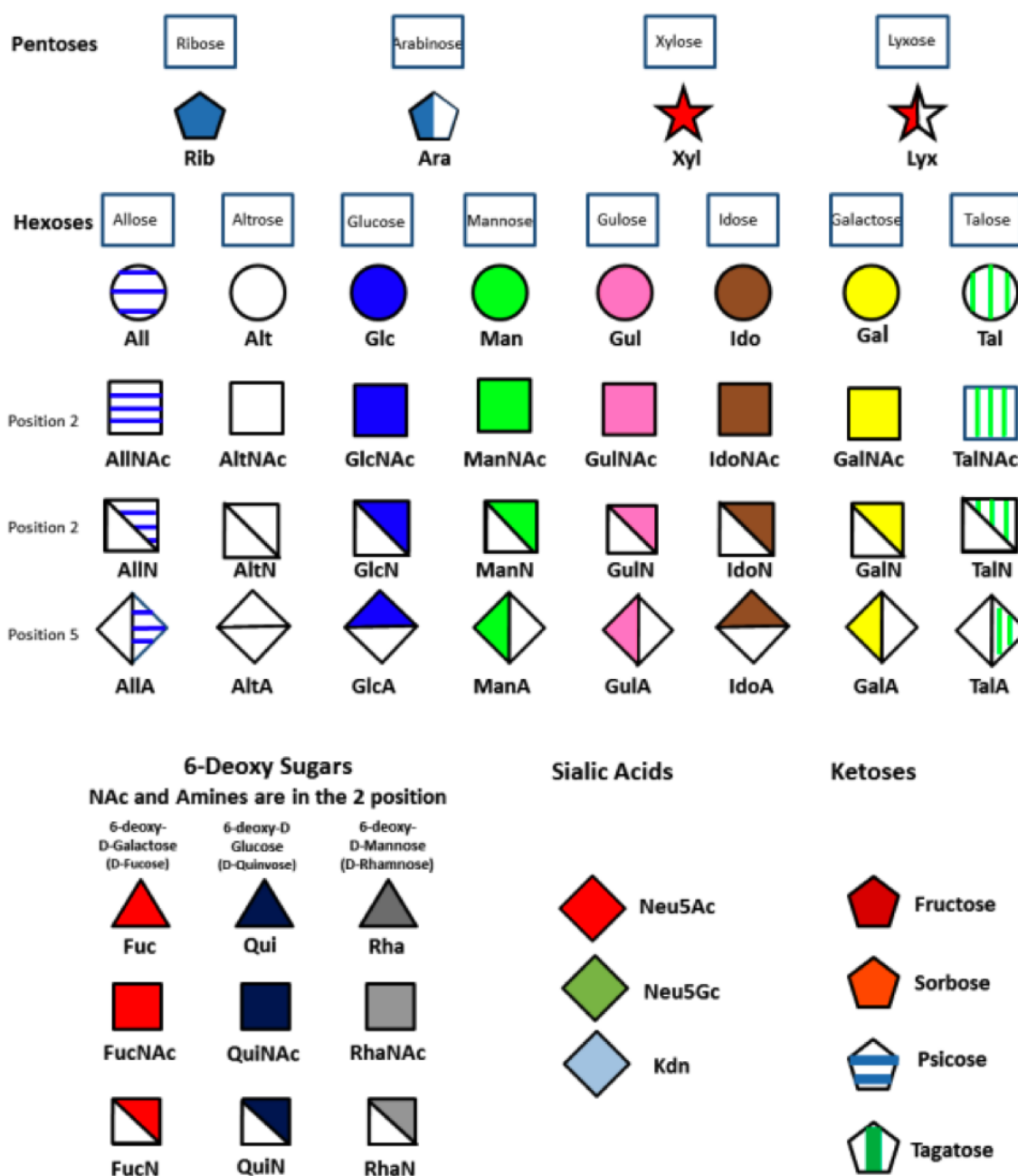


Figure 1.3: CFG Symbol Key. Reference: Public Domain <http://www.glycopedia.eu/>.

veloped over the past decade to elucidate structural characteristics of glycans and advance the knowledge of glycan functions in glycan-protein interactions. The most significant technology has been the development of glycan microarray technologies.

The glycan microarray has led to discovery of new GBPs and helped uncover novel recognition patterns in glycan interactions by known GBPs. This technology has also transformed our fundamental understanding of protein-glycan interactions across a variety of research areas including microbial pathogenesis (influenza, HIV, etc.), immunobiology, biomarker discovery, and vaccine development. The capability for high-throughput analysis and volume of data has motivated the development of computational tools to annotate, mine, and disseminate the data. The complexity of glycan structures, relative to nucleic acids and proteins, requires the development of novel approaches to expand our understanding of glycan functions and structure.

1.2.5 Bioinformatics for Glycan Microarray Data

Functional glycomics has attracted great interest due to discoveries about the important roles of complex glyco-conjugates in biological and pathological processes. Glycan microarrays of structurally defined glycans are powerful tools for functional glycomics to aid in the determination of glycan-binding specificity of GBPs. The glycan microarray was developed approximately ten years ago largely due to the Consortium for Functional Glycomics (CFG), which was funded by a multimillion dollar glue grant from the National Institute of General Medical Sciences of the NIH. The CFG was composed of several scientific cores responsible for generating novel resources and technologies for glycobiology with one, in particular, dedicated to glycan microarray screening. Core H of the CFG, the Protein-Glycan Interaction Core, analyzes investigator-generated lectins, antibodies, antisera, microorganisms, or suspected glycan-binding proteins of human, animal, and microbial origins on a mammalian glycan array to determine carbohydrate specificity and identify specific

ligands. The first printed array, developed in 2004 [15], contained 200 defined glycan targets. Over the past 10 years there has been an explosion of interest for developing glycan libraries, efficient methods of immobilization of glycans on array surfaces, and applications for analysis of glycan binding protein specificity. After several years and iterations of the glycan microarray, the current array now has 609 defined glycan targets.

In addition to glycan microarray screening, the CFG provided resources to support glycan profiling, mouse phenotyping, and gene microarrays. Each resource produced unique data sets depending on the experiment modality. The Bioinformatics Core of the CFG, Core B, was responsible for acquiring, storing, and disseminating all CFG-related data and information. Core B constructed a relational database and a website to integrate diverse data sets generated by the CFG scientific cores and participating investigators. The data is all freely available to download and explore. However, the bioinformatics platform lacks support of computational tools to perform data mining analysis on the large-scale glycomics data sets. For example, mining for glycan motifs from the glycan microarray data or tools for visualizing and mining the protein- glycan interactome.

The Resource for Informatics at Soka (RINGS) [2] is another web-based platform that that was primarily developed to provide data mining tools for large-scale glycan structure analyses. Included in this resource are kernel methods for glycan marker prediction, probabilistic models for glycan profile extraction, and frequent subtree mining of glycan structures. Missing from the RINGS platforms is the availability of advanced glycan motif analysis algorithms that utilize glycan intensity data and account for structural features that could have an impact on protein-glycan binding.

There has been considerable progress in the development of other glycan microarray platforms in addition to the CFG Glycan microarray. Having a defined file format flexible enough to support the various platforms will enable bioinformatics

tools to process experimental data across platforms in addition to facilitating the exchange of experimental data between different sites. The structural complexity of glycans coupled with various initiatives to store structural information in databases presents a challenge in glycan structure notation.

Microarrays of defined glycans represent a high throughput approach to determining the specificity of GBPs. The biological functions of glycans are exerted through their recognition by a wide variety of GBPs, which include receptors, lectins, and antibodies, as well as GBPs expressed by pathogens (viruses, bacteria, and parasites). However, the glycan determinants recognized by individual GBPs are just beginning to be understood, and is a major quest of modern glycomics research. Assessing interactions of a GBP with glycans on a microarray generates large datasets, making it difficult to identify a glycan structural motif or determinant associated with the highest apparent binding strength of the GBP. Thus, there is a clear need to automate the process for motif identification using data from glycan microarrays. A review of three of the forefront methods that have been developed in the recent years follows below.

1.2.6 Mining for Glycan Motifs on Glycan Microarray Data

Previous work on motif-based analyses [77] of binding specificities demonstrated an approach that searches for the existence of a motif from pre-defined set of 63 motifs using an intensity segregation method. In this method, motif segregation is based on fluorescence intensity, in which glycans are segregated according to the presence or absence of a given motif, and then the statistical difference in fluorescence signal is calculated between the glycans in the two groups. A drawback of this method is that it can find only a predetermined set of known motifs. This approach was refined [67] to permit the manual identification of new determinants to be included in subsequent motif segregation analysis but it is still hindered by lack of automated analysis and a

web based interface.

Compared with predefined motifs, automatic decomposition of glycan chains into substructures provides a more generalizable method for selecting binding specificity. This Quantitative Structure-Activity Method (QSAR) [115] method automatically decomposes the glycan chains into subtrees. Then, it applies Partial Least Squares (PLS) regression to the glycan array data using subtrees as features. The QSAR method uses mono-, di-, tri- and tetra-saccharide subtree features analyze binding specificities. This approach has an advantage over the Motif Segregation approach since the latter method uses pre-defined motifs. The QSAR implementation stops at tetra-saccharide subtrees, but can be extended to larger subtrees in principle. One disadvantage of this method is overlapping components. For example, a single, high-binding glycan can have multiple disaccharide components or substituents that are counted multiple times as regression coefficients. Additionally, components can be considered anywhere on the glycan but are presented or outputted without any context in terms of location (i.e. internal vs terminal). Lastly, spacer arms, of which glycans are attached in the array, are also included as part of the components.

The Glycan Miner Tool [3] is a method for mining glycan substructures based on frequent occurrence. Those substructures that appear in a sufficient number of glycan structures, bound by a minimum threshold, are considered “frequent subtrees” of the input data set. To prevent closely overlapping substructures from being returned a parameter called alpha was added to the model, defined by the following equation: $support(T') < max(alpha * support(T), minsup)$, where support (T) is the value of the support for substructure T, which is a subtree of substructure T'.

The approach is limited because it does not utilize glycan binding intensity values to determine which glycans to mine against and, similar to the QSAR method, it does not account for location of the determinant (internal vs terminal motif).

1.2.7 Deciphering and Sequencing the Human Glycome

Historically, methods for glycan sequencing were developed to address different aspects of glycan structure and included purification of glycans and application of a variety of chemistries to deduce structure. The predominant Mass Spectrometric (MS) approaches to glycan structure are limited in their ability to fully predict glycan structure that includes sequence, linkage, and anomericity. One example of high-throughput automated annotation of MS peaks was described in the Cartoonist algorithm [38], which selects annotations from a list of biologically plausible glycans based on a set of archetype cartoons manually derived from *a priori* knowledge of the biosynthetic pathways known to express certain N-Glycans. It then assigns a confidence score to the set of glycan annotations for the most abundant peaks. The annotations represent the composition and topology of the structure though portions of the structure contain specific monosaccharides and glycosidic bonds based on constraints imposed by the biosynthetic pathway.

We have yet to fully define a single meta-glycome from a source of human tissue, which indicates the challenge presented by the entire human glycome. Furthermore, no method is currently available with the precision or speed to be incorporated into an automated sequencing platform. In spite of its limitations, MS techniques, especially those that include more laborious multistage mass spectrometry have proven extremely useful in doing the deep sequencing required to fully define a glycan structure. However, a method for high throughput, deep sequencing of glycan structure will require a combination of techniques. Defined glycan microarrays provide a high-throughput approach for identifying epitopes when interrogated with GBPs that bind known glycan determinants. While the epitope alone cannot provide the information needed to fully characterize the glycan structure, it does provide a piece of “metadata” that can be applied to reveal the structure.

A method has been developed termed Metadata-Assisted Glycan Sequencing or MAGS as a structural approach that combines MS data with glycan microarray data [91, 118] and recently demonstrated its utility in defining over 20 novel structures among the human milk glycans. The glycan microarray data used in this approach needs to be obtained from libraries of relatively pure glycans, analogous to a shotgun glycan microarray where glycans, representing a selected glycome, are fluorescently derivatized, separated into relatively pure components by multi-dimensional chromatography to obtain a tagged glycan library (TGL), and printed as a shotgun glycan microarray comprising the selected glycome. It is conceivable that MAGS could be used as a general approach to define any glycome that could be presented as a shotgun glycan microarray since manual analysis of the data generated from hundreds of glycans in a microarray would not be practical.

An informatics based approach to sequencing glycans using orthogonal analytical methods can be incorporated into a single computational method that considers all the information available. This has the advantage of exploring a large sample space with the ability to constrain the set of candidate structures based on biological rules and metadata from the orthogonal methods.

1.3 Contributions of This Thesis

The complexity of glycan structures, relative to nucleic acids and proteins, requires the development of novel approaches to expand our understanding of glycan functions and structure. The central objective of this thesis is to develop novel computational methods and bioinformatics tools to advance the understanding of structure and function relationships of glycans and their recognition and binding by Glycan Binding Proteins (GBPs). The contributions of this research are described in the section below followed by a chapter of conclusions with future work.

Automated Motif Discovery from Glycan Microarray Data A challenge in the interpretation of glycan microarray data is to identify the glycan determinants that are considered necessary for binding. We present a method in [chapter 2](#) to determine the specificities of GBPs using frequent subtree mining to identify the motifs of a GBP. We evaluate this method against GBPs of known specificities to validate the results. We also demonstrate this approach revealed new recognition motifs that had not been previously reported. The work in [chapter 2](#) appears in [\[25\]](#).

Computational Approaches to Define the Human Milk Metaglycome In [chapter 3](#), we present a novel computational approach to automate the sequencing of glycans based on a method known as “Metadata Assisted Glycan Sequencing” (MAGS), which combines analyses of glycan structures by mass spectrometry (MS) and glycan microarray technology to fully characterize glycan sequences. We target the soluble glycans in the human milk glycome as the first meta-glycome to be defined using this method.

A Web-Based Bioinformatics Platform for Glycan Microarray Analysis We present our open-source web-based bioinformatics platform for glycan microarray analysis in [chapter 4](#). The platform provides interactive visualization features to view, search, and compare experimental data and also features glycan motif mining analysis. The work in [chapter 4](#) is an extension of [\[1\]](#).

Chapter 2

Automated Motif Discovery from Glycan Microarray Data

2.1 Introduction

The biological functions of glycans are exerted through their recognition by a wide variety of glycan-binding proteins (GBPs), which include receptors, lectins, and antibodies, as well as GBPs expressed by pathogens (viruses, bacteria, and parasites). However, the glycan determinants recognized by individual GBPs are just beginning to be understood [27] and is a major quest of modern glycomics research. The exploration of GBP interactions with glycans on microarrays of defined glycan structures represents a high-throughput method for exploring glycan-binding specificities [15, 32, 33, 78, 83, 94, 95, 102, 112, 120]. Such information is important in identifying potential ligands for a GBP and developing hypotheses about their roles in GBP function. For example, the observation that blood group antigens are recognized with relatively high affinity and specificity by certain human galectins, led to the discovery of their bactericidal activity and function as innate immune proteins [99].

The utility of defined glycan microarrays is dependent upon the number and diversity of glycans being interrogated with a GBP. Current glycan microarrays, such as the one publicly available from the Consortium for Functional Glycomics (CFG), contain less than a thousand glycans, which is considerably below the estimate of over 7,000 glycans/glycan determinants in the human glycomes [27]. Nevertheless, such

limited microarrays can provide immense insights into potential ligands recognized by a GBP. However, even these relatively simple analyses with a single GBP typically generate large amounts of data that are difficult to manually or visually analyze to identify glycan structural motifs or determinants required for high affinity GBP binding, as well as identify glycan substructures that interfere or preclude recognition of the glycan determinant. Thus, there is a clear need to automate the process for motif identification using data from glycan microarrays.

To address this need, we have developed an algorithm termed MotifMiner, which uses frequent subtree mining [24] to identify the motifs of a GBP. We used this algorithm to analyze data from the analyses of five biotin-labeled plant lectins and a recombinant form of human galectin-8 (Gal-8), thus representing a spectrum of binding specificities, from simple and complex glycans, respectively. Data was collected using the defined glycan microarray (version 4.0 and 4.2) from the CFG using fluorescent-labeled streptavidin for detection. In this approach the glycan microarray is interrogated with a GBP at multiple concentrations. At each GBP concentration, relative binding strength of each of the glycans on the array, related to the fluorescence intensity measured as Relative Fluorescence Units (RFU), is calculated by normalizing its RFU to a percentage of the maximum RFU for the bound glycans on the array. Non-specifically bound glycans and non-bound glycans are eliminated as binding candidates by a z-score transformation and referred to as non-binding glycans. The percentages of each binding glycan at each GBP concentration are averaged, and the data are sorted, allowing binding glycans to be ranked according to relative binding strengths. The MotifMiner algorithm then finds the common substructures in the binding glycans that exist in none or only a few non-binding glycans to identify the motifs for this GBP. Unlike other existing motif discovery methods [45, 68, 77], our method not only considers both binding and non-binding glycans, but it also can reveal unknown motifs automatically with very little user interaction.

2.2 Materials and Methods

Glycan microarray analysis The MotifMiner algorithm was developed using the CFG printed glycan microarray, v4.0 comprised of 442 glycan targets, which evolved from earlier versions with 200 glycans [15]. To determine the utility of the algorithm we used five common biotin-labeled lectins assayed at 3 different concentrations using the standard protocol for biotinylated proteins that is available on the CFG website where all raw data are located (see supplemental information). Biotinylated concanavalin A (Con A) from the jack bean (*Canavalia ensiformis*) was purchased from Vector Labs (Cat. No. B-1005, Lot No. T0829) and was assayed at 1.0, 0.1, and 0.001 $\mu\text{g}/\text{ml}$ using 5.0 $\mu\text{g}/\text{ml}$ of Cy5-labeled streptavidin (Zymed) for detection. Biotinylated Sambucus nigra agglutinin-I (SNA) was purchased from Vector Labs (Cat. No. B-1305, Lot No. T1204) and was assayed at 1.0, 0.1, and 0.01 $\mu\text{g}/\text{ml}$ using 5.0 $\mu\text{g}/\text{ml}$ of Cy5-labeled streptavidin for detection. Biotinylated peanut (*Arachis hypogaea*) agglutinin (PNA) was purchased from Vector Labs (Cat. No. B-1075, Lot No. T0918) and was assayed at 10.0, 1.0, and 0.1 $\mu\text{g}/\text{ml}$ using 5.0 $\mu\text{g}/\text{ml}$ of Alexa488-labeled streptavidin (Invitrogen) for detection. Biotinylated Ulex europaeus agglutinin-I from the Common Gorse (UEA-I) was purchased from Vector Labs (Cat. No. B-1065, Lot No. U1216) and was assayed at 30.0, 3.0, and 0.3 $\mu\text{g}/\text{ml}$ using 5.0 $\mu\text{g}/\text{ml}$ of Alexa488-labeled streptavidin for detection. Biotinylated Helix pomatia agglutinin (HPA) was purchased from Sigma (Cat. No. L6512-IMG, Lot No. 084k3776) and was assayed at 1.0, 0.1, and 0.05 $\mu\text{g}/\text{ml}$ using 5.0 $\mu\text{g}/\text{ml}$ of Alexa488-labeled streptavidin for detection. The glycan microarray slides were processed and data were collected as Excel spreadsheets as previously described [92]. The application of the algorithm to a more complex protein was done using a biotinylated preparation of Gal-8 prepared as described previously [100] and assayed at 50 and 5.0 $\mu\text{g}/\text{ml}$ using Alexa488-labeled streptavidin at 5.0 $\mu\text{g}/\text{ml}$ for detection on version 4.2

of the glycan array containing 511 glycans.

The 442 glycan targets of v4.0 are comprised of 398 unique glycans with one duplicated structure with the same linker and 43 duplicate glycans with different linkers; each is printed in replicates of six. All glycans are printed at a single concentration from a 100 μ M stock solution. Bound GBPs are detected by measuring the RFU at each of the 6 locations and averaging 4 of the 6 RFU values for each after removing the high and low values. Data are reported as average RFU in Supplementary Tables 1-6, which also include the chemical structure of the linkers associated with each glycan at the bottom of the Table. The standard deviations and %CV (100 x Std. Dev./Average) associated with each average value are reported in the original data deposited and available on the CFG website.

Ranking and selecting the binding and non-binding glycans To reveal relative binding strengths of a GBP for individual glycans, we performed analyses on the glycan array at three different concentrations of each GBP, where the selected concentrations generated signals that are in the linear range of the fluorescent scanner. The signals generated by GBP binding to specific glycans vary with concentration of the GBP based on the relative affinity of the GBP for a glycan. Before the rank is assigned to individual glycans bound by a GBP, we selected the binding glycans. To avoid using an arbitrary threshold in determining binders and non-binders, we used the z-score as the statistical test for significance of a sample. The z-score transformation is calculated by comparing the value of a sample, relative to the sample mean and standard deviation, with the null hypothesis being that a random sample pulled from the population would be a non-binder. If the converted p-value is less than 0.15, the null hypothesis is rejected and the sample is considered a binding glycan. We used a non-conservative p-value to allow more glycans in the list of candidate binders as an input to MotifMiner. The z-score transformation is based on the sum of the RFU intensity values for the three different concentrations of the glycan. This sta-

tistical test allows the program to discard not only non-binding glycans but glycans that exhibit non-specific binding, which could distort the motif discovery algorithm. These results of the statistical test are shown in Supplementary Tables 1-6 where the binding glycans are highlighted in light green. Glycans exhibiting what we define as non-specific binding generate significant binding signals, but the signals do not vary with concentration of GBP (See Supplementary Table 1, glycan #s 206, 388, 193, 203, 331, 228, 323, 216, 181, 105, 299, 413, and 427). This is in contrast with non-binding glycans, which generate low signals. In this study, we use the term non-binder or non-binding glycans to denote glycans that are either non-specific binders or non-binders.

This automatic computational method was developed to provide a systematic and objective process to define binding and non-binding glycans, and to provide an automated method for ranking glycans based on the amount of fluorescence detected, which is related to the relative affinities of the fluorescent-labeled GBP for individual glycan ligands. A rank for each binding glycan is obtained at each GBP concentration using the calculation where $\text{Rank} = 100 \times [\text{RFU Bound}/\text{highest RFU value in the assay of candidate glycans}]$. Thus, the strongest binding glycan in each assay at three different concentrations is assigned a maximal rank of 100, and the ranking values decrease with decreased binding strength. Using the three rank values for each glycan, the algorithm then calculates an average rank.

As an example of the ranking of glycans bound by a GBP, we analyzed the plant lectin SNA binding to v4.0 of the CFG glycan microarray using biotinylated SNA (Vector Labs, Burlingame, CA) at concentrations of 1.0, 0.1, and 0.01 mg/ml (Supplementary Table 1). The bound lectin was detected by the fluorescence signal from Cy5-labeled streptavidin (Zymed, Carlsbad, CA) at 0.5 mg/ml. Once ranked according to relative binding strengths, the glycans can then be inspected manually [92] or with motif analysis algorithms such as the motif discovery tool, MotifMiner,

described here or by others [65, 68, 77] for determining glycan-binding specificity or motifs for a GBP.

Motif discovery using binding and non-binding glycans The MotifMiner algorithm functions by automatically finding frequently occurring patterns in binding glycans that exist in a few or no non-binding glycans. The algorithm then incrementally discovers motifs of larger size and stops when it cannot find any motifs of the next higher size. It accomplishes this by first constructing a two-dimensional tree data structure for all the glycans from their IUPAC notation. The monosaccharides of the glycans form the nodes of the tree and their connections to other monosaccharides form the edges of the tree. A visual depiction of an example glycan tree is shown in figure 2.1, Glycan a, Step 1 where the nodes are represented by the symbols for sialic acid (Neu5Ac), fucose (Fuc), galactose (Gal), N-acetylglucosamine (GlcNAc), and mannose (Man). A monosaccharide, Man, node is highlighted with the light blue circle in Step 1, and the edges are represented with the α or β linkage(s) to the linkage position(s), 3 or 6, on the neighboring monosaccharide(s). A subtree is defined as a partial tree with a set of nodes that are attached to each other. MotifMiner initiates a search with trees having a single node. That is, each unique monosaccharide that exists in any of the binding glycans becomes a tree of size one. MotifMiner finds all the nodes that occur in a substantial number (a threshold parameter) of binding glycans. These frequent trees are expanded by adding another node if such a subtree exists in any of the binding glycans. A node (tree of size 1) or in general a subtree that is common to only a few (less than a user defined threshold, T_b) binding glycans is not expanded, as it is unlikely to be a motif. MotifMiner increases the size of the frequent common subtrees by adding another node until such an addition makes the new subtree infrequent in the binding glycans according to the set threshold. This process is shown in an example in figure 2.1 Steps 1 through 3. In step 1, Man (highlighted light blue circle) that is common to all four glycans is selected. The Gal

or Neu5Ac are not chosen as they occur in only 2 of the 4 and 1 of the 4 glycans, respectively. In step 2, the subtree is expanded to a disaccharide (highlighted in light blue) by adding another Man, and in step 3, the subtree is further expanded to the trimannosyl core (highlighted in light blue). Subtrees are expanded until they become infrequent and are not expanded further, thus generating a set of possible motifs of different sizes; there is no limit on the size of the motif that MotifMiner can discover. Once the expansion of the common subtrees among the binding glycans is completed, MotifMiner searches if any of these subtrees exist in the non-binding glycans. The subtrees that are present in only binding glycans and none of the non-binding glycans are assumed to represent motifs. The subtrees that exist in many binding glycans but may also occur in some non-binding glycans, up to a definable limit as described below, are also considered motifs. MotifMiner quickly decreases the possible motif(s) for a GBP by sorting the common subtrees according to the number of non-binding and binding glycans. Thus, the number of motifs discovered can be limited by applying a threshold parameter. The overall algorithm is described in figure 2.2.

In the application of MotifMiner to determine its utility, we used the parameter values $T_b = 4$, where T_b is the minimum number of binding glycans a subtree must exist in to be a motif; $T_n = \text{infinity}$ (essentially no threshold), where T_n is the maximum number of non-binding glycans a subtree can exist in to be a motif; and $m = 3$ where m is a parameter to limit the number of motifs returned by MotifMiner that exist both in binding and non-binding glycans. Note, all motifs that exist only in binding glycans are returned. Also, we sort the motif set in ascending order of number of non-binding glycans and then descending order of number of binding glycans. All of these settings were chosen for CFG array version 4.0 based on evaluating the results of analyses of the selected known lectins after experimenting with different values. We eliminated single nodes as motifs since single nodes (monosaccharides) are not distinguished by their positions in the tree, which would result in an excessive

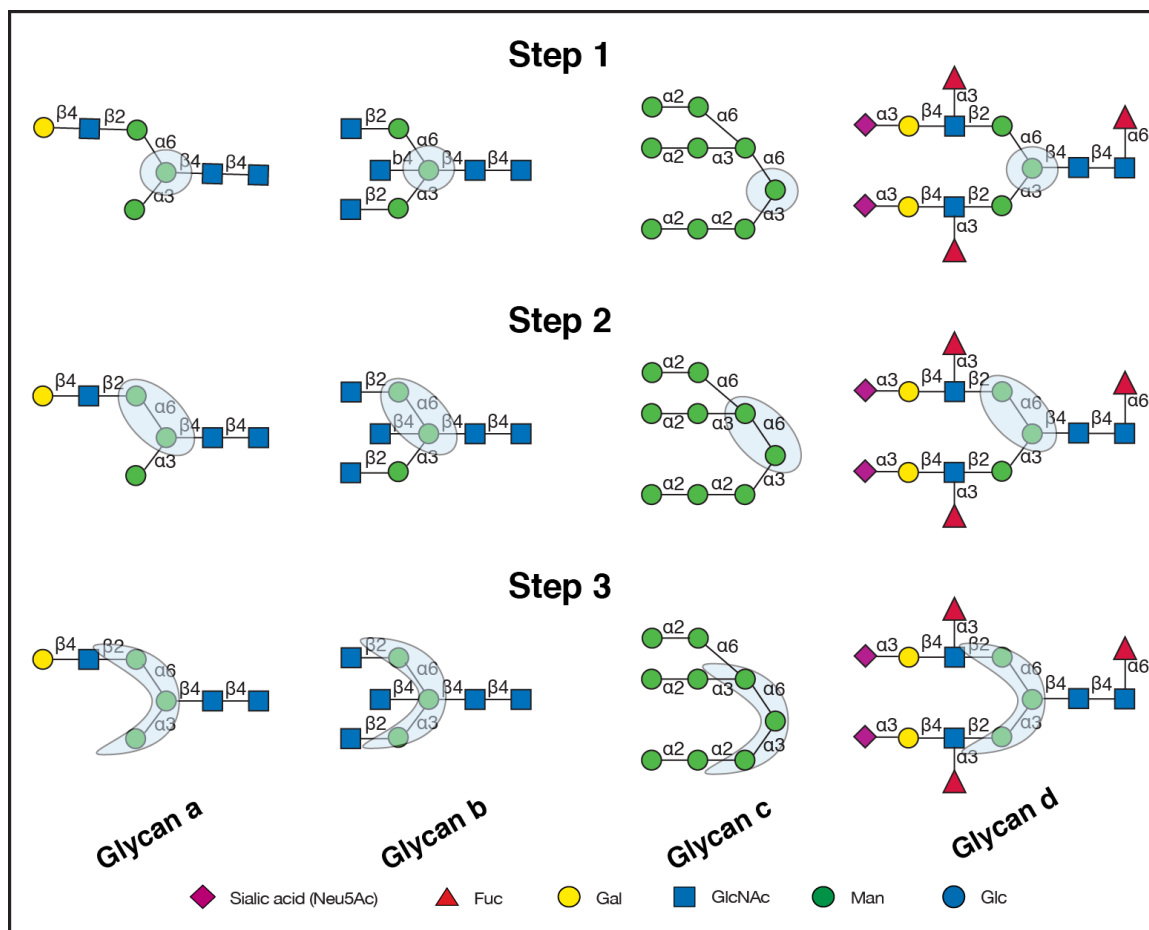


Figure 2.1: GLYMMR uses frequent subtree mining to discover the glycan-binding motifs of glycan binding proteins (GBPs). Nodes are monosaccharides represented by the symbols defined at the bottom of the figure, and their edges are represented with the a or b linkage to the linkage position on the neighboring monosaccharide. A subtree is a node or a set of nodes (highlighted in light blue). Subtrees are expanded (steps 1-3) until they become infrequent, thus generating a set of possible motifs of different sizes (Fuc, fucose; Gal, galactose; GlcNAc, N-acetylglucosamine; Man, mannose; Glc, glucose).

number of motifs. Default parameters were validated based on their ability to generate appropriate motifs for the five well-characterized plant lectins. For a different glycan microarray with larger numbers of defined structures or GBPs with more complex specificities, other parameter values might be more appropriate.

2.3 Results

To evaluate the utility of the MotifMiner program, we selected 5 biotinylated plant lectins, SNA, HPA, PNA, Con A, and UEA-I, whose glycan binding specificities are considered well-defined by a variety of methods over many years, and analyzed their binding on version 4.0 of the CFG mammalian cell glycan microarray. As an example of a more complex GBP we used recombinant, biotinylated human Gal-8, which was assayed on v4.2 of the CFG glycan microarray. The glycan microarray data were collected at three concentrations for each biotinylated lectin preparation, and the bound lectins were detected by secondary binding with either Alexa488- or Cy5-labeled streptavidin. The data were analyzed manually using the concentration dependent ranking analysis as previously described [92] and analyzed automatically with the MotifMiner program. The motifs discovered for each of the 5 plant lectins are represented in figure 2.3 6-10 in symbol format where each motif, if it occurs as a glycan on the array, is identified by its glycan number and ranking. In addition, each motif has associated descriptors indicating the number of glycans on the array containing the motif that were binding glycans (binders), and the number of glycans containing the motif on the array that were non-binding glycans (non-binders). Inspection of the structural differences between non-binding and binding glycans possessing the motif(s) provides information of the specificity of the GBP. The more complex motif analysis of Gal-8, a protein with two unrelated carbohydrate-binding domains, is shown in figure 2.10.

MotifMiner Algorithm

1. Initialize each unique node among all the strong binding glycans as a subtree of size 1.
Let this set be S.
2. For each subtree in S:
 - o Calculate the number of strong binding glycans containing S.
 - o If the subtree occurs in more than T_s (threshold parameter) strong binding glycans then:
 - add it to a set of expandable subtrees (ES), and also to a set of possible motifs (PM).
3. If the set ES is not empty:
 - o Create an empty set NewS.
 - o For each subtree in ES
 - Expand the subtree by adding a node such that the new subtree exists in at least one of the strong binding glycans and add these subtrees to a set NewS.
 - o S = NewS
 - o Go to step 2
4. For each subtree in the set of possible motifs PM:
 - o Count the number of strong binding glycans (n_s) and weak binding glycans (n_w) in each PM.
 - o If $n_s > T_s$ and $n_w < T_w$, where T_s and T_w are threshold parameters, add the subtree to a set of motifs, M.
5. Sort the set of motifs, M, in descending order of the number of strong binding glycans (n_s) containing M and in ascending order of the number of weak binding glycans (n_w) containing M.
6. From the sorted set M, output all the motifs that exist only in strong binding glycans from the sorted set and m motifs that exist in both strong and weak binding glycans, where m is a parameter to limit the number of results. Optionally, eliminate motifs that are substructures of larger motifs to reduce the similar motif structures.

Figure 2.2: Summary of GLYMMR algorithm. GLYMMR uses a repetitive interrogation of increasingly larger subtrees to discover motifs.

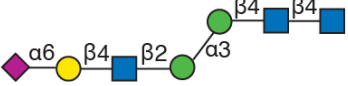
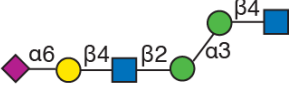
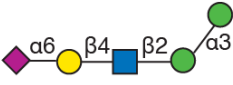
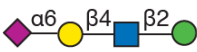


SNA Motif	Motif Structure	Number of Strong Binders	Number of Weak Binders	Glycan ID Number	Average Ranking
a		9	0	343	71 %
b		9	0	NA	NA
c		9	0	NA	NA
d		13	1	NA	NA
e		21	1	257, 258	46,51 %
f		22	4	263	0 %

Figure 2.3: Display of motifs for SNA. The structures of motifs (a-f) discovered for SNA over three concentrations are indicated using symbols defined in figure 2.2 with a and b anomeric carbons and linkage positions to the adjacent monosaccharides indicated. The glycan ID number indicating the position on v4.0 of the CFG glycan microarray (Supplementary Table 1; see online supplementary material at <http://www.liebertpub.com>) is shown for motifs found as glycans on the array with its corresponding average ranking calculated by the algorithm in parentheses. Motifs discovered by the algorithm that are not found as glycans on the array have no glycan ID or ranking and are designated NA (not applicable). The number of bound glycans containing motif is determined by the algorithm and indicates the number of bound glycans found on the glycan array that contain the corresponding motif, while the number of non-bound glycans containing motif indicates the number of glycans found on the glycan array that contain the motif, but are considered non-binding glycans by the algorithm. The display of the graphical user interface used to generate this summary is shown in Supplementary Figure S1 (see online supplementary material at <http://www.liebertpub.com>).

Analysis of SNA The results of the motif analysis generated for SNA using data in Supplementary Table 1 and with the default algorithm settings (no filter, $m=3$ and normal sorting as described in figure 2.2) is shown in figure 2.3. Results with different parameter settings are shown in Supplementary Fig. 1, which provides the display of the web-based graphical user interface generated by MotifMiner. The motifs a, b, and c are the primary motifs that occur in 9 binding glycans, but do not exist in any non-binding glycans. Motifs d, e, and f exist in some non-binding glycans along with many binding glycans. According to our parameter, m , only three motifs are shown that exist in both binding and non-binding glycans. Note that MotifMiner discovered motifs of different sizes as it searched frequent subtrees of increasing size until the subtrees are infrequent in binding glycans. All of the motifs discovered by the algorithm are related in that they possess the minimal motif f (Neu5Ac α 2-6Gal). The web-based graphical user interface allows the investigator to inspect the structures of the non-binding glycans and binding glycans possessing a motif by “clicking” on the underlined number next to binding glycans (binders) and non-binding glycans (non-binders) associated with the motif (See Supplementary Fig. 1a).

For example, there are twenty-two binding glycans (not shown) that contain motif f (Neu5Ac α 2-6Gal) and four non-binding glycans (figure 2.3 and Supplementary Fig. 1a). The structures of the four non-binding glycans, possessing the minimal motif f (glycan # 263), are displayed by clicking on the number 4 associated with the “non-binders” for that structure and viewing the resulting graphic user interface (Supplementary Fig. 2), which is summarized in figure 2.4. The results indicate that the lack of SNA binding by these glycans is consistent with the results of manual inspection analysis of SNA using the identical data [92]. Inspection of the motifs discovered by MotifMiner indicate that the lectin prefers the determinant Neu5Ac α 2-6Gal β 1-4GlcNAc (motif e, # 257, 258) at a ranking of 46% and 51% (figure 2.3) compared to the determinants Neu5Ac α 2-6Gal β 1-4Glc (#261, 262) at a ranking of 0%

(figure 2.4) or Neu5Ac α 2-6Gal β (motif f, #263) also at 0% (figure figs. 2.3 and 2.4). The non-binding of N-glycan #313 (figure 2.4) containing the preferred trisaccharide, Neu5Ac α 2-6Gal β 1-4GlcNAc (motif e) at a ranking of 0% is due to the fact that this trisaccharide determinant is located on the 6-branched mannose of the biantennary glycan, while a Neu5Ac α 2-3Gal β 1-4GlcNAc, a non-determinant is on the 3-branched mannose of #313. This observation is consistent with our manual analysis that led to the conclusions that SNA binds to the trisaccharide determinant Neu5Ac α 2-6Gal β 1-4GlcNAc when it is present on the 3-branched mannose of the biantennary structure [92], as found in glycan #343 and in motifs a, b, and c of figure 2.3, but will not bind to the trisaccharide determinant when it is present on the 6-branched mannose of the biantennary structure. The motif, Neu5Ac α 2-6Gal β 1-4GlcNAc (#257 and 258), is also a component of the four larger motifs and is found in twenty-one binding glycans (not shown) and in only one non-binding glycan (#313, figure 2.4) where it is found on the 6-branched mannose of the biantennary structure mentioned above. Thus, motif e (figure 2.3), which is found in motifs a-d, represents a minimal motif, since glycan #263 (motif f in figure 2.3) and Neu5Ac α 2-6Gal β 1-4Glc (#261 and 262, figure 2.4) are not bound (0%). Interestingly the disaccharide Neu5Ac α 2-6Gal (#263), in spite of being a commonly recognized determinant for SNA and being necessary for binding, is not sufficient alone for SNA binding. The specificity of SNA determined by the MotifMiner was consistent with the conclusions based on manual inspection of the data in Supplementary Table 1 as described previously [92], but using this program deciphering this specificity can be accomplished in a fraction of the time.

A graphical description of the discriminatory capability of MotifMiner is illustrated in figure 2.5, in which the subsets of all glycans are exhibited relative to their recognition by SNA. Twenty-two glycans of the total four hundred and forty-two in the glycan microarray possess the determinant Neu5Ac α 2-6Gal β 1-4GlcNAc. MotifMiner reported that 21 of them that possessed that motif were bound by SNA and

identified one as a non-binding glycan. Similar types of graphical illustration showing the selectivity of each lectin for subsets of glycans can be easily generated from datasets described below.

Analysis of HPA HPA is a lectin isolated from the albumin gland of the edible snail *Helix pomatia*, and specifically agglutinates human blood group A, but not B and O erythrocytes [44, 104]. This historically-defined specificity results from its binding a group of defined glycans with the following order of preference: Forssman antigen (GalNAc α 1-3GalNAc-R), blood group A substance (GalNAc α 1-3(Fuc α 1-2)Gal β 1-R), Tn antigen (GalNAc α -Ser/Thr), GalNAc, and GlcNAc [113]. For this reason, HPA has been classified as an α -GalNAc-binding lectin and used extensively to define the presence of this sugar residue at the non-reducing termini of mammalian oligosaccharides. The results of the motif analysis generated for HPA using glycan array data on v4.0 of the CFG array (Supplementary Table 2) with the algorithm set with no filter, m=3 and normal sorting (see supplemental data showing the web-based graphical user interface in Supplementary Fig. 3) are summarized in figure 2.6 where two very different motifs were discovered. Motifs a and b are related and clearly consistent with the known specificity for terminal α -GalNAc, but motif c terminates in α -GlcNAc. This second glycan recognition by HPA has been largely overlooked, since α -GlcNAc occurs rarely in mammals and has so far been found only in O-glycans of human gastric mucins [73, 119]. HPA was recently described as a member of a group of invertebrate lectins that are involved in innate immunity in the snail, where it participates in the protection of fertilized eggs by agglutinating a variety of microorganisms [87]. Unlike mammalian cells, microorganisms may be more likely to present α -GlcNAc on their surfaces.

Inspection of the non-binding and binding glycans associated with the α -GalNAc terminating sequences (motifs a and b in figure 2.6) indicates that binding by HPA is associated with terminal α -GalNAc residues, as previously described with one ex-

ception. That exception is the presence of a GalNAc at the reducing end of a glycan expressing a difucosylated blood group A-like sequence as seen in $\text{GalNAc}\alpha 1-3(\text{Fuc}\alpha 1-2)\text{Gal}\beta 1-4(\text{Fuc}\alpha 1-3)\text{GlcNAc}\beta 1-3\text{GalNAc}$ (#415), as shown in Supplementary Table 2. HPA does not recognize this glycan, although it binds well to glycan #80 with the simpler sequence $\text{GalNAc}\alpha 1-3(\text{Fuc}\alpha 1-2)\text{Gal}\beta 1-4(\text{Fuc}\alpha 1-3)\text{GlcNAc}$. The structural difference between #80 and #415 that confers such a remarkable difference in HPA recognition is not known, but was easily revealed using the motif discovery algorithm.

Inspection of the non-binding and binding determinants associated with the other HPA binding motif *c*, $\text{GlcNAc}\alpha 1-4\text{Gal}$ (figure 2.6), shows that this determinant is present in the five binding glycans (glycans #s 333, 334, 335, 336, and 338 in Supplementary Table 2, red highlight) and in two non-binding glycans (#s 337 and 339 at 21% and 12%, respectively, Supplementary Table 2). The non-binding glycan #337 is closely related to the five binding glycans sharing the terminal sequence, $\text{GlcNAc}\alpha 1-4\text{Gal}\beta 1-4\text{GlcNAc}\beta 1-3\text{Gal}$, with the binding glycan #s 333, 336, and 338. This is the result of not meeting the criteria for statistical significance after performing the z-score transformation due to the lower than anticipated binding at the highest HPA concentration. While the data in Supplementary Table 2 show a ranking value of 21 for glycan 337, the automated calculation considered it a non-binding glycan. Glycan #339 at 12% also fell below the cutoff, but this reflects a significant difference in structure, which possesses a reducing terminal GalNAc (#339, $\text{GlcNAc}\alpha 1-4\text{Gal}\beta 1-3\text{GalNAc}$). Simple inspection of the motif analysis results provides subtle specificity information indicating that HPA must be affected in its binding by more than the presence of the terminal monosaccharide, since a modification of the structure $\text{GlcNAc}\alpha 1-4\text{Gal}\beta 1-4\text{GlcNAc}$ - (#335) to $\text{GlcNAc}\alpha 1-4\text{Gal}\beta 1-3\text{GalNAc}$ - (#339) has a significant negative impact on HPA recognition (Supplementary Table 2).

Analysis of PNA The lectin from peanut (*Arachis hypogaea*) agglutinates

SNA Motif	Motif Structure	Number of Strong Binders	Number of Weak Binders	Glycan ID Number	Average Ranking
f		22	4	263	0 %

SNA Motif f Weak Binders	Glycan ID Number	Average Ranking
	263	8 %
	261, 262	0, 0 %
	263	0 %

Figure 2.4: Motif f of SNA discovered in figure 2.3 occurs in bound and non-binding glycans on v4.0 of the microarray. Motif f (Neu5Aca2-6Gal) is found in 22 bound glycans and in only 4 non-binding glycans. The 4 non-binding glycans that contain motif f are indicated using symbols defined in figure 2.2, with a and b anomeric carbons and linkage positions to the adjacent monosaccharides indicated, and with the glycan ID number indicating the positions on v4.0 of the CFG glycan microarray (Supplementary Table S1; see online supplementary material at <http://www.liebertpub.com>), and their corresponding average rankings. The display of the graphical user interface used to generate this summary is shown in Supplementary Figure S2; see online supplementary material at <http://www.liebertpub.com>).

neuraminidase-treated human erythrocytes [14, 103] and was originally designated “anti-T agglutinin” because, like the mammalian anti-T antibody, it induced T-polyagglutination that was associated with certain bacterial and viral infections [62]. The T-antigen was shown to have the structure Gal β 1-3GalNAc-R [98] and purified peanut agglutinin (PNA) was shown to be very specific for this disaccharide sequence, which is typically found within core 1 and core 2 O-glycans. The proposed binding site on the protein was restricted to the non-reducing terminal β -linked Gal of the T-antigen [62].

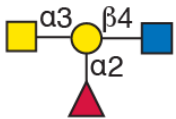
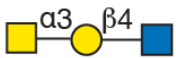
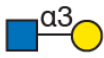
HPA Motif	Motif Structure	Number of Strong Binders	Number of Weak Binders	Glycan ID Number	Average Ranking
a		8	4	343	71 %
b		8	4	NA	NA
c		5	2	NA	NA

Figure 2.5: Discriminatory capability of GLYMMR. The subsets of all glycans are displayed based on their structural attributes relative to their recognition by SNA. Twenty-two glycans of the total 442 on v4.0 of the CFG microarray are recognized as binding glycans by SNA, and all of those glycans contain the sequence Neu5Aca2-6Galb1-4GlcNAc.

The results of the motif analysis generated for PNA using glycan array data on v4.0 of the CFG array (Supplementary Table 3) and with the algorithm set with no filter, $m=3$ and normal sorting (see supplemental data showing the web-based graphical user interface in Supplemental Fig. 4) is summarized in figure 2.7. The disaccharide Gal β 1-3GalNAc-R (motif c) was among the three motifs identified by this algorithm. The largest motif was GlcNAc β 1-6(Gal β 1-3)GalNAc (motif a), which

is a composite of the other two disaccharide motifs, GlcNAc β 1-6GalNAc (motif b) and Gal β 1-3GalNAc (motif c). Thus, the motif analysis discovered the T-antigen (glycans #131, 133 in Supplementary Table 3, and motif c in figure 2.7) among the motifs for this lectin, but the trisaccharide, GlcNAc β 1-6(Gal β 1-3)GalNAc (motif a; glycan #s 125 and 182 ranked 80% and 75%, respectively in Supplementary Table 3) was a stronger binder (higher ranking) than the simple T-antigen disaccharide alone. Although motif b (glycans #183, and 184 in Supplementary Table 3) itself is non-binding glycan, the algorithm includes this motif since it appears in the binding glycans. Inspection of the ranking of the glycans bound by PNA (Supplementary Table 3) indicated that the core 2 O-glycans tetrasaccharide, Gal β 1-4GlcNAc β 1-6(Gal β 1-3)GalNAc (glycan #s 157 and 159 ranked 98% and 89%, respectively), pentasaccharide, Gal β 1-3GalNAc α 1-3(Fuc α 1-2)Gal β 1-4GlcNAc (glycan #377 ranked 85%) and the trisaccharide, GlcNAc β 1-6(Gal β 1-3)GalNAc (motif a; glycan #s 125 and 182 ranked 80% and 75%, respectively); all are ranked higher than the disaccharide, Gal β 1-3GalNAc (glycan #s 131 and 133 ranked 57% and 73%, respectively). This is another example of how increased numbers of glycans available on glycan microarrays provide new paradigms for lectin specificities. The apparent strict specificity and the proposed restricted binding site on PNA for the T-antigen disaccharide must clearly be revised based on the fact that many diverse glycans exist that bind PNA stronger than the simple T-antigen disaccharide found within O-glycans. It is noteworthy that PNA did not recognize any glycan in which the disaccharide motif Gal β 1-3GalNAc was sialylated or fucosylated.

Analysis of Con A The MotifMiner analyses of the lectins described above were based on data generated after the glycans were ranked and averaged over three concentrations. During this process each data set is submitted to analysis by the ranking program utilizing the z-score transformation, and the ranking of each glycan at three concentrations is averaged and the glycans are sorted by rank from high to low

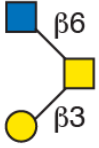
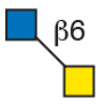

PNA Motif	Motif Structure	Number of Strong Binders	Number of Weak Binders	Glycan ID Number	Average Ranking
a		7	3	125, 182	80, 75 %
b		7	10	183, 184	0, 0%
c		18	27	131, 133	57, 73 %

Figure 2.6: Display of motifs for HPA. The structures of motifs (a-c) discovered for HPA over three concentrations are indicated using symbols defined in figure 2.2, with a and b anomeric carbons and linkage positions to the adjacent monosaccharides indicated. The glycan ID number indicating the position on v4.0 of the CFG glycan microarray (Supplementary Table S2; see online supplementary material at <http://www.liebertpub.com>) is shown for the motif found as a glycan on the array with its corresponding average rank in parentheses. Motifs discovered by the algorithm that are not found as glycans on the array have no ID number or ranking and are designated NA (not applicable). The number of bound glycans containing motif is determined by the algorithm and indicates the number of bound glycans found on the glycan array that contain the corresponding motif; while the number of non-bound glycans containing motif indicates the number of glycans found on the glycan array that contain the motif but are considered non-binding glycans by the algorithm. The display of the graphical user interface used to generate this summary is shown in Supplementary Figure S3; see online supplementary material at <http://www.liebertpub.com>).

for motif discovery analysis. The ranking program and MotifMiner were designed to analyze three concentrations of GBP, however, these analyses can also be performed on one or two data sets. If multiple concentrations are selected for analysis they should be in a linear range of the fluorescence scanner used for the analysis (between 1,000 and 40,000 to 60,000 RFU). If data are included that are not in this linear range, the motif pattern will be altered based on the bias generated by the non-linear data. This effect can be demonstrated in the analysis of Con A from the jack bean (*Canavalia ensiformis*), known to be specific for either terminal α -linked mannose or glucose residues on glycans [39, 40] and for the internal trimannosyl core structure of N-glycans [9, 26, 71]. The biotinylated lectin was assayed at 1.0, 0.01 and 0.001 $\mu\text{g}/\text{ml}$ and detected with Cy5-labeled Streptavidin. The data are shown in Supplementary Table 4b, where the RFU values for the 50 strongest binding glycans at 1.0 $\mu\text{g}/\text{ml}$ were between 30,000 and 60,000 RFU, and z-score transformation resulted in 57 binding glycans that were statistically significant. At the highest concentration of the lectin, the ranking is difficult due to the inability to discriminate between binding and non-bound glycans.

The motif analysis generated for Con A with the algorithm set with no filter, $m=3$ and normal sorting (see supplemental data showing the web-based graphical user interface in Supplementary Fig. 5), based on the all three concentrations where the high concentration is well above the linear range is summarized in figure 2.8. The motif analysis using these data discovers the well-characterized terminal α -mannose specificity [40] of Con A in motif a (figure 2.8) and a portion of a biantennary N-glycan (motifs b-d), which is known to be a motif of Con A [9, 71], but none of the motifs discovered occur as an individual glycan on the array. However, if the motif analysis is carried out using only the two low concentrations of Con A, which are clearly in the linear range of the instrument, the motif with the same algorithm settings (web-based graphical user interface Supplementary Fig. 6) is more representative of what is known



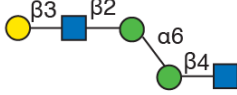
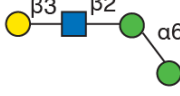
Con A Motif	Motif Structure	Number of Strong Binders	Number of Weak Binders	Glycan ID Number	Average Ranking
a		10	0	NA	NA
b		11	2	NA	NA
c		11	2	NA	NA
d		11	2	NA	NA

Figure 2.7: Display of motifs for concanavalin A (Con A) based on three concentrations of the lectin. The structures of motifs discovered for Con A (a-d) are indicated using symbols defined in figure 2.2, with a and b anomeric carbons and linkage positions to the adjacent monosaccharides indicated. The glycan ID number indicating the position on v4.0 of the CFG glycan microarray (Supplementary Table S4b; see online supplementary material at <http://www.liebertpub.com>), is shown for the motif found as a glycan on the array with its corresponding average rank in parentheses. Motifs discovered by the algorithm that are not found as glycans on the array have no ID number or ranking and are designated NA (not applicable). The number of bound glycans containing motif is determined by the algorithm and indicates the number of bound glycans found on the glycan array that contain the corresponding motif, while the number of non-bound glycans containing motif indicates the number of glycans found on the glycan array that contain the motif, but are considered non-binding glycans by the algorithm. The display of the graphical user interface used to generate this summary is shown in Supplementary Figure S5; see online supplementary material at <http://www.liebertpub.com>).

for this lectin and is summarized in figure 2.9. Con A was initially characterized as a mannan-binding lectin with specificity for terminal $\text{Man}\alpha 1-2$ [39, 40], found in motif a of both figure 2.8 and figure 2.9. This motif a does not appear on the array as a glycan, but it is present in ten highest ranked mannans as shown in Supplementary Table 4a,b, while no non-binding glycans possess this determinant. As mentioned above, Con A is also known to bind biantennary N-glycans that possess the trimannosyl core structure shown in motif b, which appears as glycan #204 on the array ranked at 88%. Motifs c and d (figure 2.9) both contain motif b and this trimannosyl core with the core GlcNAc (motif c) or core chitobiose (motif d; glycan #s 47 and 48 at rankings of 53% and 66%, respectively) and are found in thirteen binding glycans and fifty-four non-binding glycans. Of the fifty-four non-binding glycans, many are triantennary, bisected, and derivatized N-glycans that possess this chitobiose, trimannosyl core structure, which are known to be poorly bound by Con A [9]. These observations clearly demonstrate the need to work with data that are in the linear range of the analysis to obtain representative motifs.

Analysis of UEA-I UEA-I is a lectin isolated from *Ulex europaeus* (Common Gorse), and specifically binds H (Type 2) glycans that represent the O blood group antigen [65]. The known specificity of UEA-I for blood group H (Type 2) is clearly presented as the motif b (figure 2.10) discovered by the MotifMiner program using data from three concentrations of this lectin as shown in Supplementary Table 5 with the algorithm set with no filter, $m=3$ and normal sorting (see supplemental data showing the web-based graphical user interface in Supplementary Fig. 7). The H (Type 2) structure, $\text{Fu}\alpha 1-2\text{Gal}\beta 1-4\text{GlcNAc}$ (motif b), includes the other discovered motifs a and c, both of which are presented as glycan targets on the array (glycan #s 75 and 160, respectively). Consistent with early observations on this lectin [10, 29, 76, 101], the $\alpha 1-2$ -linked fucose is an absolute requirement for UEA-I binding. This is based on the observations that the non-fucosylated $\text{Gal}\beta 1-4\text{GlcNAc}$ (LacNAc), which

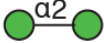

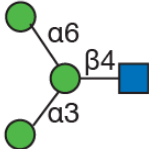
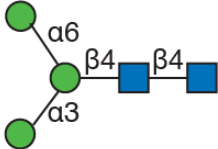
Con A Motif	Motif Structure	Number of Strong Binders	Number of Weak Binders	Glycan ID Number	Average Ranking
a		10	0	NA	NA
b		19	54	204	88 %
c		13	54	NA	NA
d		13	54	47, 78	53, 66 %

Figure 2.8: Display of motifs for Con A based on two concentrations of the lectin. The structures of motifs discovered for Con A (a-d) are indicated using symbols defined in figure 2.1, with a and b anomeric carbons and linkage positions to the adjacent monosaccharides indicated. The glycan ID number indicating the position on v4.0 of the CFG glycan microarray (Supplementary Table S4a; see online supplementary material at <http://www.liebertpub.com>), is shown for the motif as a glycan on the array with its corresponding average rank in parentheses. Motifs discovered by the algorithm that are not found as glycans on the array have no ID number or ranking and are designated NA (not applicable). The number of bound glycans containing motif is determined by the algorithm and indicates the number of bound glycans found on the glycan array that contain the corresponding motif, while the number of non-bound glycans containing motif indicates the number of glycans found on the glycan array that contain the motif, but are considered non-binding glycans by the algorithm. The display of the graphical user interface used to generate this summary is shown in Supplementary Figure S6; see online supplementary material at <http://www.liebertpub.com>).

is a component of most of the binding glycans (Supplementary Table 5), is found in over one hundred-eighty non-binding glycans, lacks fucose addition, and is not bound by UEA-I.

Motif b is present at two locations on the array with slightly different spacers as glycan #s 72 and 73 ranked at 70% and 79%, respectively. Higher-ranking glycans were mono- and di-sulfated derivatives of fucosylated LacNAc or lactose (Supplementary Table 5), which have not been previously reported as inhibitors of UEA-I binding, and were not identified as a motif since there were only three sulfated structures on the array and the algorithm was set for $T_b = 4$. Interestingly, the Lewis y (Ley) antigenic tetrasaccharide determinant $[\text{Fuc}\alpha 1-2\text{Gal}\beta 1-4(\text{Fuc}\alpha 1-3)\text{GlcNAc}]$ was found in two of the eleven binding glycans containing motif b as glycan #s 68 and 69 ranked at 39% and 41%, respectively. The lower ranking relative to motif b suggests that the additional fucose on the Ley-related structure destabilizes its binding to UEA-I, which is consistent with previous studies on inhibition of blood group substance precipitation by UEA-I [101]. Again, although the Ley-containing glycans are found among the binding glycans, MotifMiner does not identify this as a motif since it exists in only two bound glycans and the threshold, T_b , in our algorithm was set to 4. Motif b also occurs in twenty-eight non-binding glycans, of which thirteen were related to blood group A, $\text{GalNAc}\alpha 1-3(\text{Fuc}\alpha 1-2)\text{Gal}\beta 1-4\text{GlcNAc}$, and six were related to blood group B, $\text{Gal}\alpha 1-3(\text{Fuc}\alpha 1-2)\text{Gal}\beta 1-4\text{GlcNAc}$, indicating that substituting the Gal of motif b with anything at its 3-position destroyed UEA-I binding, which is consistent with the H(O) specificity [65] of UEA-I; however, substitution of this motif at the 4-position (glycan #s 90 with terminal $\alpha 4\text{GalNAc}$ and #112 with $\alpha 4\text{Gal}$, ranked at 58% and 74%; Supplementary Table 5) did not significantly affect binding. The weaker-binding Ley tetrasaccharide is ranked even lower when it is linked $\beta 1-3$ to GalNAc at the reducing end (#413), $\beta 1-2$ to Man in di- and tri-antennary N-glycans (#358, #442), or $\beta 1-3$ to Gal in longer fucosylated polylactosamines (#66 and #67).


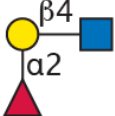
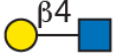
UEA-1 Motif	Motif Structure	Number of Strong Binders	Number of Weak Binders	Glycan ID Number	Average Ranking
a		16	62	75	13 %
b		11	28	72, 73	70, 79 %
c		11	183	160	0 %

Figure 2.9: Display of motifs for UEA-I. The structures of motifs (a-c) discovered for UEA-I over three concentrations are indicated using symbols defined in figure 2.1, with a and b anomeric carbons and linkage positions to the adjacent monosaccharides indicated. The glycan ID number indicating the position on v4.0 of the CFG glycan microarray (Supplementary Table S5; see online supplementary material at <http://www.liebertpub.com>) is shown for the motif found as a glycan on the array with its corresponding average rank in parentheses. Motifs discovered by the algorithm that are not found as glycans on the array have no ID number or ranking and are designated NA (not applicable). The number of bound glycans containing motif is determined by the algorithm and indicates the number of bound glycans found on the glycan array that contain the corresponding motif, while the number of non-bound glycans containing motif indicates the number of glycans found on the glycan array that contain the motif but are considered non-binding glycans by the algorithm. The display of the graphical user interface used to generate this summary is shown in Supplementary Figure S7 (see online supplementary material at <http://www.liebertpub.com>).

Thus, the binding pattern of UEA-I on v4.0 of the CFG array demonstrates exquisite specificity for H (Type 2) glycans with complete lack of binding to blood group A, B, and H (Type 1) glycans as previously described [76, 101], and differential ranking of the known glycan structures provides a definition of the UEA-I specificity. The MotifMiner program permits a rapid identification of the binding motif and facilitates deciphering the specificity by providing the structures of the binding and non-binding glycans containing each motif.

Analysis of Recombinant, Human Galectin-8 (Gal-8) Human Gal-8, a member of the galectin family of galactose-binding proteins, was selected for analysis using MotifMiner as an example of a biologically relevant and complex glycan-binding protein. Gal-8, a tandem repeat galectin, whose single polypeptide has two carbohydrate recognition domains (CRDs) that bind different sets of glycans [20, 100], was assayed at two different concentrations and analyzed using MotifMiner. As expected for a GBP with mixed specificity, Gal-8 binds strongly to 46 glycans (Supplementary Table 6) on the CFG glycan array v4.2 with the strongest being #162, LNnT ($\text{Gal}\beta 1-4\text{GlcNAc}\beta 1-3\text{Gal}\beta 1-4\text{Glc}$) at a rank of 86%, while 20 binding glycans express human blood group A or B. Glycans containing poly-N-acetylglucosamine sequences ($\text{Gal}\beta 1-4\text{GlcNAc}$)_n comprised a significant number of the binding glycans and several sialylated and sulfated glycans were among this group. These data are consistent with previous analyses of Gal-8 on the CFG array [100]. When the data from the two concentrations (50 $\mu\text{g}/\text{ml}$ and 5.0 $\mu\text{g}/\text{ml}$) were analyzed using MotifMiner with the default settings of $m=3$, normal sorting and no filtering, 5 motifs were discovered as shown in the display of the web-based graphical user interface in Supplementary Fig. 8a. With these settings the algorithm discovered an extended Blood group A determinant [$\text{Gal}\beta 1-3\text{GalNAc}\alpha 1-3(\text{Fuc}\alpha 1-2)\text{Gal}$], the blood group A [$\text{GalNAc}\alpha 1-3(\text{Fuc}\alpha 1-2)\text{Gal}$] and LNnT ($\text{Gal}\beta 1-4\text{GlcNAc}\beta 1-3\text{Gal}\beta 1-4\text{Glc}$), but did not discover the blood group B motif [$\text{Gal}\alpha 1-3(\text{Fuc}\alpha 1-2)\text{Gal}$]. By increasing the value of m to 9,

we were able to discover the Blood Group B motif as shown in Supplementary Fig. 8b, which shows 11 motifs, that can be filtered to remove subsets of identical structures to show the 7 motifs in Supplementary Fig. 8c and summarized in figure 2.10. The inability of the algorithm to detect blood group B motifs is due to the fact that the parameter setting of 'm=3' restricted the number of motifs that exist both in binding and non-binding glycans to 3. Increasing this value resulted in the discovery of the human blood group B motifs. In addition, the array contains only four Gal-8 binding glycans containing the blood group B tetrasaccharide determinant, Gal β 1-3(Fuc α 1-2)Gal β 1-4GlcNAc, while there are seven Gal-8 binding glycans containing the blood group A tetrasaccharide determinant, GalNAc β 1-3(Fuc α 1-2)Gal β 1-4GlcNAc. Thus, MotifMiner discovered both the human blood group motif, as well as the poly-N-acetyllactosamine motif of Gal-8, which is known to possess different specificities due to its two different CRDs, and by modifying the parameters of the algorithm it is possible to accommodate the discovery of multiple motifs in a single sample.

2.4 Discussion

Our results show that MotifMiner is a glycan binding motif discovery algorithm that first ranks and classifies glycans interrogated with a GBP into binding and non-binding glycans and then identifies the motifs based on the glycan structures. The motif discovery results are shown for 5 representative and well characterized lectins SNA, HPA, PNA, Con A, and UEA-I. Each motif discovered by MotifMiner was consistent with the published and accepted specificities of these well known plant lectins that have been used for many years for detecting specific glycan structures, although our findings expand the knowledge about the glycan motifs recognized by several of these lectins. The results of MotifMiner can be presented to the user on a web-based graphical user interface (Supplementary Figs. 1-8) for quick interpretation in a web-based format.

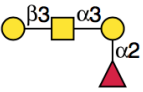
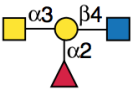


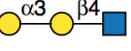
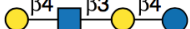

Galectin-8 Motif	Motif Structure	Number of Bound Glycans Containing Motif	Number of Non-Bound Glycans Containing Motif	Glycan ID Number (Average Rank)
a		5	0	NA
b		12	2	84 (40%), 85 (60%)
c		18	6	NA
d		7	12	NA
e		5	10	NA
f		5	3	161 (47%), 162 (86%)
g		5	7	180 (30%)

Figure 2.10: Display of motifs for human recombinant Gal-8. The structures of motifs (a-g) discovered for Gal-8 over two concentrations are indicated using symbols defined in figure 2.1, with a and b anomeric carbons and linkage positions to the adjacent monosaccharides indicated. The glycan ID numbers indicating the position on v4.2 of the CFG glycan microarray (Supplementary Table S6; see online supplementary material at <http://www.liebertpub.com>) are shown for motifs found as glycans on the array with their corresponding average ranks in parentheses. Motifs discovered by the algorithms that are not found as glycans on the array are designated NA (not applicable). The number of bound glycans containing motif is determined by the algorithm and indicates the number of bound glycans found on the glycan array that contain the corresponding motif, while the number of non-bound glycans containing motif indicates the number of glycans found on the glycan array that contain the motif but are considered non-binding glycans by the algorithm. The display of the graphical user interface used to generate this summary is shown in Supplementary Figure S8, panel c (see online supplementary material at <http://www.liebertpub.com>).

MotifMiner offers several advantages to finding glycan binding motifs using microarray approaches. In contrast to the algorithm described here, other methods for motif discovery either limit the motifs to a predefined list or focus on only binding glycans while ignoring the presence of the motif in non-binding glycans. The work by Porter et al. [77] demonstrated an approach that searches for the existence of a motif from pre-defined set of 63 motifs using an intensity segregation method. In this method, motif segregation is based on fluorescence intensity, where glycans are segregated according to the presence or absence of a given motif, and then the statistical difference in fluorescence signal is calculated between the glycans in the two groups. A drawback of this method is that it can find only a predetermined set of known motifs. This approach was recently refined [68] to permit the manual identification of new determinants to be included in subsequent motif segregation analysis. The other available method for motif discovery [45] provides a definition of the common subtrees among a list of glycans without evaluating binding intensities.

The MotifMiner algorithm coupled with a web-based graphical user interface to evaluate results addresses the limitations in prior methodology. MotifMiner is a relatively simple system for exploring the CFG glycan microarray for the purpose of identifying the glycan-binding specificity of GBPs. It incorporates a non-biased selection of candidate glycans, based on their binding strength, with a simple approach to searching for glycan determinants or motifs and reports both the binding and non-binding glycans in which the determinants or motifs occur. While the algorithm itself does not generate the specificity for a GBP, it provides the researcher with the ability to quickly find and compare the structural features of the binding and non-binding glycans so that specificity determinations can be quickly made.

In evaluating the utility of MotifMiner by comparing the known specificities of a selection of plant and animal lectins, we found the algorithm to be useful and accurate if applied to sound datasets with an understanding of the limitation of the program

based on the threshold parameters used. We suggest a default setting for the threshold parameters for our initial analyses so that the variables would be associated with the data used for input as we developed the application with the 442 glycan targets of v4.0 of the CFG array. As we discovered motifs of known lectins, we found that modifying the method for the value of m was useful in revealing all of the known motifs. The modifications were useful as the compliment of structures on the array can bias the sorting parameters. Filtering the results so that subsets are combined into the largest glycan limited the number of motifs. It may be necessary to make further modification of parameters such as the sorting method as the application is applied to later versions of the array; i.e., v5.0 comprised of 611 glycan targets. The application to newer versions of the array requires only the addition of the structures to the library of glycans, and it is possible to apply the algorithm to any array of defined structures.

This algorithm permits a rapid and accurate method for defining subsets of glycans when it is necessary to work with large numbers of structures that have to be represented with a relatively complex nomenclature. For example the data generated to support the analysis of the lectin SNA can be described in subsets according to the diagram in figure 2.5. There are 442 glycan targets printed on v4.0 of the CFG array. For SNA, which is considered to be a sialic acid-specific lectin, there are two major subsets of glycans; i.e., sialylated structures comprising a subset of 93 glycans and non-sialylated glycans making up a subset of the remaining 349. The subsets within the sialylated glycans can be subdivided first into those with a terminal sialic acid coupled to Gal by all possible linkages (Sia-Gal-R), which make up a subset of 81 glycans, that is further reduced to a subset of 60 by eliminating all except those with the terminal trisaccharide Sia-Gal-GlcNAc in all possible linkages, which would include both type 1 (Gal β 1-3GlcNAc) and type 2 (Gal β 1-4GlcNAc) glycans, as well as some glycans having modified sialic acids. Restricting that subset to those containing only

Neu5Ac α 2-6Gal β 1-4GlcNAc results in 22 glycans. MotifMiner reported 21 glycans containing Neu5Ac α 2-6Gal β 1-4GlcNAc or motif a in figure 2.3. The other glycan on the array that contains motif a is glycan #313, which was considered a non-binder by the algorithm as shown in figure 2.4. Similar subset distributions are easily carried out for all of the lectins analyzed.

MotifMiner successfully defined motifs of all of the well-characterized lectins that we selected for analysis. We had previously shown by manual inspection of data from SNA binding in the CFG glycan array that this lectin, which was generally thought to be specific for Neu5Ac α 2-6Gal, prefers that this disaccharide be on a Type 2 glycan sequence (Neu5Ac α 2-6Gal β 1-4GlcNAc), for which it exhibits stronger binding than to sialylated lactose, which lacks the GlcNAc residue and contains Glc instead [92]. These data are consistent with the original studies of Goldstein and colleagues showing that SNA binds glycans with the sequence Neu5Ac α 2-6Gal(NAc)-R [90]. Interestingly, MotifMiner revealed that SNA binds well to the Neu5Ac α 2-6Gal β 1-4GlcNAc determinant within N-glycans except when it is linked to the α 6-branched mannose of an N-glycan [92]; the motifs discovered by MotifMiner were all related to the heptasaccharide (motif a shown in figure 2.3). The ability to compare the structural features of binding and non-binding motif-containing glycans, provided a rapid method for confirming what aspects of a glycan structure destabilize binding. For example, the presence of a non-binding determinant on the 3-branch of glycan number #313 (figure 2.4) strongly supported the conclusion that SNA will recognize the determinant Neu5Ac α 2-6Gal β 1-4GlcNAc on the 3-branch as a binding motif, but not if the determinant is on the 6-branched mannose.

The analysis of HPA resulted in the discovery of two extremely different motifs, terminal α -linked-GalNAc and α -linked-GlcNAc, and provided an interesting example of how the numbers of glycans on a microarray and the threshold parameters of MotifMiner interact. As shown in figure 2.6, for HPA binding there are eight binding

glycans that possess terminal α -linked-GalNAc, which is consistent with prior studies on its binding specificity [113]. However, MotifMiner also identified five binding glycans possessing terminal α -linked-GlcNAc, which is a relatively rare glycan in animals [73, 119], and not previously described to be well recognized by HPA. If glycans with this latter structural feature had not been on the array, this motif would not have been discovered. In fact, if it had been represented less than 4 times on the array, it would have been identified as a binding glycan, but not as a motif, since the threshold parameter, T_s , was set at 4. Thus, when using MotifMiner, it is important to initiate the specificity determination by listing all of the binding glycans sorted from high to low ranking in order to identify any rare glycan that is a binding glycan for the GBP in question. For example, if one inspects the ranking of glycans that are bound by SNA (Supplementary Table 1), the glycan #s 353 ($\text{Kdn}\alpha 2\text{-6Gal}\beta 1\text{-4GlcNAc}$) and 256 ($\text{Neu5Ac}\alpha 2\text{-6Gal}\beta 1\text{-4(6OSO3)GlcNAc}$) are the highest ranking glycans; this suggests the possibility that SNA might prefer sulfated glycans or other sialic acid derivatives, but these motifs do not arise from analysis by MotifMiner, since they are represented less than four times on the array. Interestingly, in other work on arrays comprised of derivatives of sialic acid, we observed that SNA preferred Kdn and its acetylated derivatives to Neu5Ac and Neu5Gc, the more common mammalian sialic acids [96]. These results also illustrate the need in glycan microarray studies for presenting glycan determinants in a variety of formats and backbones on multiple glycans to facilitate identification of motifs.

MotifMiner revealed new motifs associated with PNA binding that had not been previously reported. The discovery of new motifs for well-defined GBPs will probably become common as MotifMiner and other algorithms are applied to additional well-defined lectins. This will occur largely because glycan microarrays allow us to evaluate hundreds of glycans simultaneously as opposed to the classical method of hapten inhibition of lectin binding that was generally accomplished using a limited

numbers of structures and a single oligosaccharide per experiment. With hundreds of glycans on a single array, unexpected discoveries become common. The original studies defining the specificity of PNA as being directed against Gal β 1-3GalNAc are certainly valid and the definition of PNA as an anti-T agglutinin is appropriate [62], but the fact that more extended structures are also equivalently recognized, such as the non-sialylated core 2 O-glycan Gal β 1-4GlcNAc β 1-6(Gal β 1-3)GalNAc, should now be taken into account. The utility of using multiple concentrations for analysis of GBP specificities is demonstrated in data obtained for PNA, as shown in Supplementary Table 3 where the glycans #206, 299, 203 and 150 are in bold type. These glycans received a low but significant ranking of 3-7% in this analysis; however, these glycans, which obviously do not contain a PNA motif, received high rankings when lower concentrations of lectin were analyzed due to the lower value of the maximum RFU. When averaged, this high ranking at low lectin concentration resulted in an elevated average ranking. These glycans are examples of non-specific binders since the RFU values do not change as a function of concentration, and they are eliminated from the candidate glycans by the z-score transformation described above. Because MotifMiner eliminates the non-binding glycans (non-specifically bound and non-bound) prior to calculating their rank, the average rankings generated by the algorithm (figure figs. 2.3, 2.4 and 2.6 to 2.10) may be different from the average rankings generated manually in the Supplementary Tables 1-6.

We purposely evaluated Con A, which binds mannose-containing N-glycans [40], at a concentration outside the range of linearity to demonstrate the effects of such data on binding discrimination using MotifMiner. Using non-linear data results in a bias where weaker binding glycans receive a higher than appropriate rank, which caused the algorithm, with the set parameters, to miss the trimannosyl core as a motif (figure 2.8). When we eliminated the non-linear data, the analysis Motif Miner discovered the trimannosyl core as a motif (figure 2.9) demonstrating the requirement

of high quality data and the ability of the algorithm to use less than three data sets for analysis. In fact, the analysis can be used with a single data set if the data are in a linear range; however, identifying non-specific binding glycans, which do not vary with concentration of GBP would not be identified.

The analysis of UEA-I revealed the well-know specificity of this lectin for the H Type 2 structure [65], but some unexpected observations are made. In this case the inspection of the ranking of glycans (Supplementary Table 5) indicates that the three highest-ranking glycans are sulfated derivatives of the UEA-I determinant (glycan #s 251, 213, 212), which would not be detected as motifs because they were not found on a sufficient number of glycans. These data suggest that sulfated glycans may be important in natural UEA-I-glycan interactions.

We included the analysis of Gal-8 on a different version of the array (figure 2.10) to demonstrate the limitations of a single set of parameters and the versatility of the algorithm to perform well using any glycan array that is entered into the database with diverse and complex lectins. While the initial set of parameters worked well for GBPs with a single Carbohydrate Recognition Domain (CRD) like Con A, SNA, HPA, PNA, and UEA-I, the evaluation of GBPs with multiple specificities or multiple CRDs required an expanded display of motifs and the ability to filter out motifs that are substructures of larger motifs. The user can change these parameters easily using web-based interface for any lectin analysis without the need to modify the parameters in algorithm itself. This would give the users the flexibility to investigate both simple and complex lectins in even larger arrays while adjusting the number of motif structures they want to view depending on if they want to do a quick or thorough analysis. As future work, we will add the ability to change other parameters such as Tb (the minimum number of binding glycans a subtree must exist in to be a motif) and Tn (maximum number of non-binding glycans a subtree can exist in to be a motif) from the user interface. Also, we will develop more filtering options like eliminating all the

smaller motifs if they are part of a larger motif irrespective of their binding glycans.

In summary, MotifMiner is a useful motif discovery algorithm that is incorporated into a web-based application. It has been tested and is publicly available at <http://motifminer.emory.edu> for versions 4.0 and 4.2 of the CFG mammalian cell glycan microarray with the data presented in this manuscript, and it will be expanded for use on all versions of the CFG glycan microarray. Participating Investigators of the CFG and the public can use it to define the specificity of GBPs, which are difficult to accomplish by manual inspection of hundreds of different glycan structures.

Chapter 3

Computational Approaches to Define the Human Milk Metaglycome

3.1 Introduction

Glycans play integral roles in many essential biological functions including cell signaling, molecular recognition, immunity, and inflammation generally via their specific interactions with proteins [108]. Unlike the template driven process for synthesizing linear nucleic acids and proteins, glycans are enzymatically synthesized and are thus products of many genes forming linear and branched sequences of stereospecific monosaccharides that provide unique surfaces for protein interactions [11]. Understanding the specificity of glycan-binding proteins (GBPs) provides clues to their functions, and defining specificity is accomplished by comparing the structures of glycans bound by a GBP with related structures that are unbound [92]. This requires that the glycan ligands be completely defined, which is not a trivial task. In addition, while the human glycome is estimated to be at least 10 times larger than the proteome [27]; to date, no glycome or meta-glycome (partial or sub-glycome) related to a tissue, organ or cell type has been defined. Furthermore no method is currently available with the requisite precision or speed to be incorporated into an automated sequencing platform.

Historically, methods for glycan sequencing were developed to address different aspects of glycan structure and included purification of glycans and application of a variety of chemistries to deduce structure [72]. The predominant Mass Spec-

trometric (MS) approaches to glycan structure are limited in their ability to fully identify glycan structure that includes sequence, linkage, and anomericity; however, a large amount of information is generated from mixtures of glycans. One example of high-throughput automated annotation of MS peaks was described in the Cartoonist algorithm [38]. The algorithm initially selects annotations from a list of biologically plausible glycans based on a set of archetype cartoons manually derived from apriori knowledge of the biosynthetic pathways known to express certain N-glycans. It then assigns a confidence score to the set of glycan annotations for the most abundant signals. The annotations represent the composition and topology of the structure though portions of the structure contain specific monosaccharides and glycosidic bonds based on constraints imposed by the biosynthetic pathway. A de novo approach, termed STAT [34], predicts glycan structures by generating all possible topologies of a glycan structure based on the precursor ion mass, charge carrier, and product ion mass from the MSn data. A review [110] of automated interpretation of MS Spectra for glycan structures provides an overview of various approaches; however, all of the existing approaches fall short of fully characterizing glycan structures including the linkage and anomericity. In spite of its limitations, MS techniques, especially those that include more laborious multistage MS [6] have proven extremely useful in doing the deep sequencing required to fully define a glycan structure. However, a method for high throughput, deep sequencing of glycan structure will likely require a combination of techniques. Defined glycan microarrays provide a high-throughput approach for identifying epitopes on individual glycans when interrogated with GBPs that bind known glycan determinants. While the determinant or epitope alone cannot provide the information needed to fully characterize the glycan structure, it does provide a piece of “metadata” that can be applied to reveal the structure. We developed Metadata-Assisted Glycan Sequencing or MAGS as a structural approach that combines MS data with glycan microarray data [91, 118], and recently demonstrated its utility in

defining over 20 novel structures among the human milk glycans [7, 117]. The glycan microarray data used in this approach is acquired from libraries of relatively pure glycans, analogous to a shotgun glycan microarray [93] where glycans, representing a selected glycome, are fluorescently derivatized, and separated by multidimensional chromatography to resolve isomeric structures and obtain relatively pure components in a tagged glycan library (TGL). The separated glycans are then printed as a shotgun glycan microarray comprising the selected glycome. We reasoned that MAGS could be used as a general approach to define any glycome that could be presented as a shotgun glycan microarray; however, manual analysis of the data generated from hundreds of glycans in a microarray would be a tedious process. We therefore developed a software package, termed GlycomeSeq, to sequence human milk glycans (HMGs) through automated meta-analysis of experimental data based on our MAGS [91] approach.

We reasoned that a key to more automatable sequencing is to create a “virtual glycome” comprising all theoretical structures, and then test the predicted structures from the obtained metadata against this virtual glycome; we also used a novel algorithm to filter candidate glycan structures through this knowledge base to arrive at a single structure consistent with all available information. For this study we selected the free glycans of human milk, since the glycans are easily accessible and a large literature is available about them. In addition to the nutritional disaccharide lactose (10-15 g/l), human milk contains a complex mixture of larger, free oligosaccharides or glycans (5-10 g/l) that are not efficiently metabolized in the stomach of the neonate and reach the intestines where they are thought to have probiotic activity [17], as well as to provide protection against pathogens by interfering with adhesion of pathogens to intestinal epithelial cells as “decoy receptors” [50, 57, 85, 117]. Milk glycans have also been implicated in having beneficial innate immune and immunomodulatory effects [30], decreasing colon contractility [12], and promoting gut epithelial cell mat-

uration [48, 57, 85]. Recent studies on the biological functions of human milk have indicated that breast fed infants have soluble milk glycans circulating in blood at detectable levels suggesting potential systemic effects of such glycans [37]. The structures and quantities of the free glycans in human milk vary widely among individual mothers based on their genetics, which controls the expression of the human Lewis blood groups and time of lactation [17]. A variety of factors influence the repertoire of the human glycome, including genetics, environment, and time (or conditions of synthesis); and like the human genome, the glycome of the human milk free glycans is an “average set of structures” that represents all of the possible structures that could exist at any one time or in any one individual. Thus, there is great interest in defining the human milk metaglycome and its components, since they may differ in many ways between different sources and the overall repertoire of glycans in a milk sample is subject to many variables. The computational approach developed here was successful in identifying glycan structures within the human milk metaglycome and the approach should be applicable to other cellular and tissue metaglycomes.

3.2 Methods

Collection of metadata from a tagged glycan Library and its corresponding shotgun glycan microarray In prior studies related to the current development we developed a functional glycomics approach using a shotgun glycan microarray (SMG) of human milk [117, 118], in which the SGMs are interrogated with viruses and antibodies that recognize unique glycan determinants. We also used defined lectins and antibodies before and after specific exoglycosidases to obtain detailed information about the repertoire of glycan determinants in individual glycans within the library. These data generated large quantities of metadata on each glycan; we then used logic to arrive at a structural solution based on the identification of specific determinants by antibody and lectin binding [117]. However, the manual data pro-

cessing to arrive at structures was extremely time consuming, and we reasoned that an algorithm could be used to apply the logic generated from the structural information and specificity of glycosyltransferases available from previous studies on milk glycans. If this could be accomplished as an automated, high throughput method, it could be applicable as a general approach to defining a human metaglycome.

To address the human milk meta-glycome and to develop an automated system for metadata-assisted glycan sequencing (MAGS), we used data available from the SGM from 10 different donors with mixed blood groups [117], and selected data from the analysis of 42 human milk glycans and 14 standards (Supplemental Table 1). This included 33 glycans whose structures were predicted by manual analysis of MALDI-TOF data, antibody and lectin binding data, and MSn analysis [7, 117]. In this paper, we describe the approach to developing an automated system for MAGS, and we show how automated analysis of multiple modalities can enrich the set of predicted structures for an unknown glycan target by introducing the concept of a “virtual glycome” of human milk free glycans as a knowledge base and an algorithm for filtering candidate structures from the virtual glycome.

Generating the virtual human milk soluble glycome Based on previous studies of human milk soluble glycan structures and the specificities of enzymes involved in their synthesis, we established a set of biologically plausible rules that can be used to define all of the possible structures synthesized as free glycans in human milk without regard to differential genetics or stage of lactation (See Supplement A). These rules can then be used to computationally generate and store all possible glycan structures including structural isomers into a database. Seeking to establish a general method to generate a virtual glycome, we developed a novel approach using regular expressions to represent the biosynthetic rules for a particular metaglycome. For our initial attempt, we focused solely on the human milk soluble glycome and we describe the method below.

The Virtual Glycome Generator algorithm is initialized with the core Lactose structure, a threshold parameter for the maximum core size of the glycome, and the patterns that represent the extensions and terminal modifications of HMG biosynthesis. The actual representation of the extension and terminal modification patterns are declared using regular expressions as described in Supplemental Figure 1. A Regular Expression (RE) is a sequence of symbols or characters (also known as a string) that represent a set of patterns that describe regular languages from Formal Language Theory. Common applications of REs include validating email addresses in a web form, replacing or extracting values in a text file, UNIX shell commands such as `ls` or `grep`, etc. REs can use operators and meta-characters to express non-trivial patterns of strings including repeating characters, optional characters, and classes of characters. In our Virtual Glycome Generator algorithm, we are interested in the class of regular languages that are star-free, such that the language described is finite but also supports the union, disjunction, and finite repetition of patterns.

A RE is typically used to check for a match between an input string and a pattern. In our case, we are interested in generating all the input strings that can possibly represent a pattern. This turns our problem into the question of how can we generate all the possible strings (or glycan structures) that match a given regular expression? From the field of Automata Theory, we know we can use a Finite-State Automaton (FSA) to implement a regular expression. An FSA recognizer starts in an initial state and transitions to other states based on the sequence of symbols from the input, and if the automaton is in an accepting state when the input terminates, then it is said to accept the input, thus representing a match against the regular expression.

Since we are actually interested in generating output from the FSA, we reasoned we could simulate the automaton to output all possible strings that would be accepted by the FSA. We used a non-recursive backtracking algorithm to find all possible ac-

cepting states. The algorithm accumulates the input characters for each transition until reaching an accepting state; when it outputs the resulting string to an array. We used the Linear Code nomenclature [31] to represent the set of symbols for the monosaccharides, anomericity, linkages, and branches to define the regular expressions. We describe the full algorithm to generate the virtual glycome in Supplement B.

The set of glycan structures generated by this algorithm represents the “virtual glycome” for human milk free glycans and is used as the knowledge base for glycan structure prediction. Since glycans commonly form branched structures, we utilized Extensible Markup Language (XML) as the format to store generated glycan structures in the virtual glycome database. XML has the advantage of supporting hierarchical queries using the XML Path Language (XPath). XPath is analogous to a SQL query used to query records from a relational database though XPath provides a mechanism to query and navigate through hierarchical representations of data, a feature that is well suited for branched glycan structures and utilized extensively in the candidate glycan filtering step described in the Glycan Structure Prediction method.

Glycan Structure Prediction The GlycomeSeq algorithm requires a spreadsheet input file (Supplement C) that contains the composition and glycan array binding data for each glycan target. The composition of the glycans are reported based on the numbers of residues of hexose (H), which represents the single reducing terminal Glc plus Gal, which is the only other hexose found in human milk free glycans; N-acetylhexosamine (N), which is GlcNAc; deoxyhexose (F), which is Fucose; and Sialic Acid (S), which is only Neu5Ac. For example, the composition of the sialyl Lewis a epitope, $\text{Neu5Ac}\alpha\text{2-3Gal}\beta\text{1-3(Fuc}\alpha\text{1-4)GlcNAc}\beta\text{1-3Gal}\beta\text{1-4Glc}$, is H3N1F1S1; and the core structure (without Fucose and Sialic acid additions) would be H3N1. We proceed through the steps of the algorithm below.

As illustrated in figure 3.1, the algorithm initially selects all structures from the

database that match the composition of the unknown target. We then evaluate the positive binders and select the intersection of candidate structures from the remaining set of structures that contain the determinant for the positive binding GBPs. The selection of the candidate structures that contain the determinant is based on an XPath query defined for the GBP. Each of the XPath queries is defined in Supplement C.

Next, for each non-binding GBP, we filter out candidate structures that contain the determinant for each non-binding GBP using an XPath query. This step is completed for all negative binders to eliminate candidate structures for that target. Finally, if we have additional determinants as identified by MSn reporter ions (e.g. linear lactose, branched lactose, terminal fucosylated LacNAc, etc.) we select the intersection of candidate structures in our final set of predicted structures.

3.3 Results

The virtual human milk soluble glycan glycome In our initial study we have focused only on the neutral free glycans of human milk. We use our Virtual Glycome Generator to store the virtual human milk soluble glycan glycome into a database. Table 1 shows the theoretical number of glycans that can exist for each composition based on the biosynthetic rules of human milk glycans (Supplement A) up to a dodecaose core structure (H7N5). While core structures (without fucose or sialic acid) as large as H10N8 have been reported, the amounts of glycans with core structures greater than H5N3 are vanishingly small. Nevertheless, the virtual glycome of human milk soluble glycans is certainly greater than 50,000 different possible structures. The web-based software tool, GlycomeSeq, provides the number of isomers within each composition and will display the structures of all isomers within each composition (<https://glycomeseq.emory.edu/>). In an effort to further our understanding of the functional and structural roles of human milk glycans we registered acces-

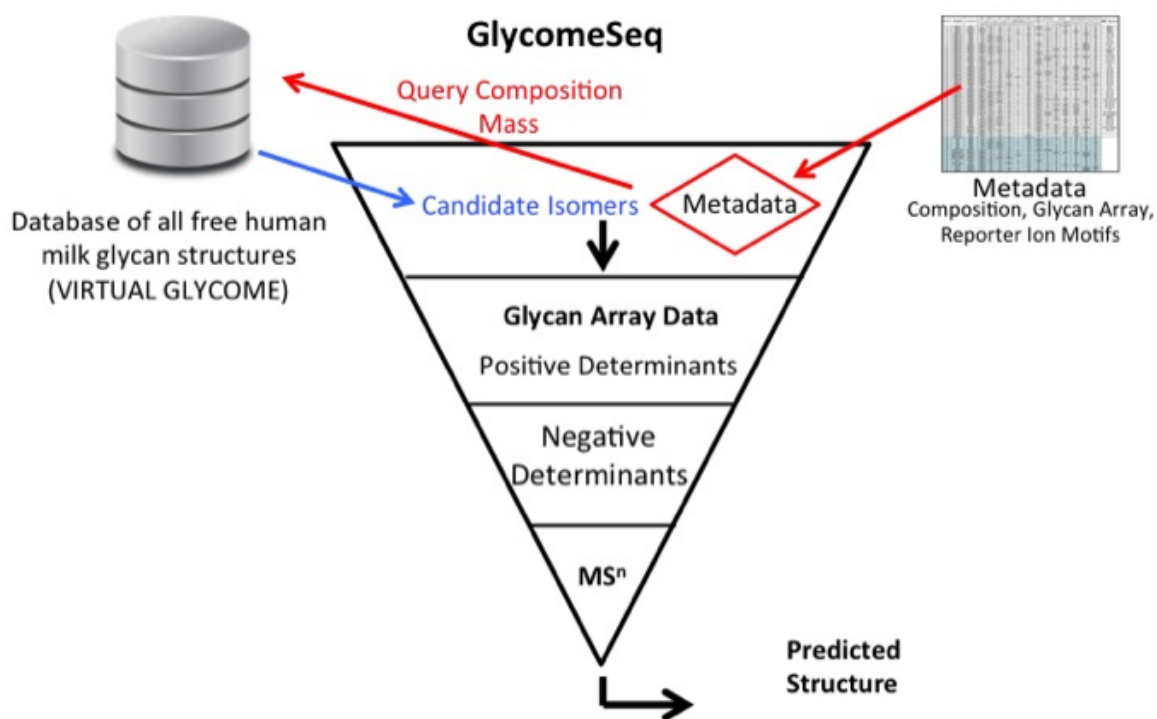


Figure 3.1: GlycomeSeq Algorithm. Combines the metadata from MS and glycan array binding data to select and filter out candidate structures from the virtual glycome.

sion numbers in GlyTouCan <http://glytoucan.org> for the predicted structures that were verified by independent structural methods. We also shared our virtual HMG database with UniCarbKB [19], which can be found at <http://unicarbk.org/milk>. The UniCarbKB platform is a knowledgebase that allows public access to a curated database of glycan structures and associated metadata including publications and glycan structural classification by taxonomy, tissue, protein, etc. In addition to sharing the glycan structures, GlycomeSeq also provides link outs to the UniCarbKB website for each of the predicted structures where the user can view any associated metadata for the glycan structure.

Identification of determinants within the structures of human milk glycans

The simple compilation of glycan structures is considered to have limited value since little information on function can be generated from lists of structures. However, having a compilation of all of the possible structures within a particular glycome may permit some useful predictions regarding the relationship of structure and function. While the total number of possible free milk glycans with disaccharide to dodecasaccharide core structures is estimated to be 53,514 (Table 1), the region of a complex carbohydrate molecule that is required for the specific recognition of a biologically relevant GBP has been termed the glycan determinant [27] which is comprised of di- to pentasaccharides; the number of determinants is significantly less than the total number of structures. As an exercise to test this hypothesis, we identified the number of non-reducing, terminal determinants from di- to pentasaccharide determinants in the human milk metaglycome as a function of increasing core structure size. Examples of determinants identified by the defined lectins and antibodies used in this study are shown in Supplemental Fig. 2. The results of this analysis are shown in figure 3.3.

While the total number of structures in each core size increases dramatically as shown by the number of glycans in each core size (figure 3.3, parentheses), the number

Sialic acid and Fucose	Core Structures (no fucose or Sialic acid)						
	H2	H2N1	H3N1	H4N2	H5N3	H6N4	H7N5
F0S0	1	0	2	4	10	26	72
F1S0	2	0	5	12	41	135	454
F2S0	1	0	4	13	66	291	1229
F3S0	0	0	1	6	52	335	1860
F4S0	0	0	0	1	20	220	1715
F5S0	0	0	0	0	3	81	982
F0S1	2	0	3	8	23	71	230
F1S1	2	0	7	22	91	358	1420
F2S1	0	0	5	21	140	745	3751
F3S1	0	0	1	8	104	822	5517
F4S1	0	0	0	1	37	513	4920
F5S1	0	0	0	0	5	178	2710
F0S2	0	0	1	5	19	75	299
F1S2	0	0	2	12	71	363	1794
F2S2	0	0	1	9	101	717	4578
F3S2	0	0	0	2	67	739	6454
F4S2	0	0	0	0	20	421	5460
F5S2	0	0	0	0	2	129	2814
Total glycans in each core	8	0	32	124	872	6219	46259
Cumulative Total	8	8	40	164	1036	7255	53514
	Disaccharide core	Triose core	Tetraose core	Hexaose Core	Octaose Core	Decaose Core	Dodecaose Core

Figure 3.2: The Virtual Glycome of Human Milk Free Glycans. Core Structures, which are unsubstituted with fucose or sialic acid as indicated by the designation F0S0 in the top row, are designated by composition where H represents hexose (a single reducing terminal glucose and galactose residues) followed by the number of residues; i.e., H2 is the lactose core structure (Gal β 1-4Glc), and N represents GlcNAc residues i.e., H3N1 represents the isomers comprised of a single reducing terminal glucose, 2 galactose residues, and a single GlcNAc. The composition H2N1 was included to be comprehensive, but this structure is not found as a free glycan in human milk (See Supplement A). The numbers for each composition indicate the number of isomers with the indicated composition that can be biosynthetically generated in human milk based on the rules described in Supplement A. For example, the composition H5N3F2S1 is shared among 140 isomeric structures in the virtual glycome.

of terminal determinants increases to an apparent asymptotic value; i.e., 13 terminal disaccharide determinants for all human milk free glycans, 21 terminal trisaccharide determinants, 43 terminal tetrasaccharide determinants, and 89 terminal pentasaccharide determinants. Thus, while the number of possible free glycan structures in any human milk sample may be enormous, there are a limited number of potentially relevant biologically active determinants that we would predict to be recognized by GBPs and other glycan recognition molecules. Interestingly, over 90% of the determinants in each determinant size are represented in the glycans with a core composition of H5N3, and glycans with larger core structures are found in vanishingly small amounts. These observations address the questions of how many possible free glycans can exist in human milk and how many of these structures may be biologically relevant. Although the number of possible structures of human milk glycans may be well over 50,000 and depend on individual genetics, time since initiation of lactation, diet, and possible time of day, the number of biologically relevant determinants may be less than 100. The free glycans of human milk may, therefore, present a cluster or bouquet of thousands individual structures where these biosynthetic pathways may have evolved to support a relatively limited number of biologically relevant determinants. By this unique biological process the microheterogeneity within a metaglycome is less important than the total number of relevant determinants expressed.

GlycomeSeq Algorithm The GlycomeSeq algorithm has been implemented in a software package that executes within a web application (<http://glycomeseq.emory.edu>). It takes an input file in a format described in Supplement C; however the example on the website is preloaded with the data from Supplement C and the web application is available at <http://glycomeseq.emory.edu>. The data used for demonstrating our automated sequencing approach was obtained from the validation of the HM-SGM reported previously [117]. The data from 11 GBPs with known specificities binding to 42 purified human milk glycans and 14 standards of known structure

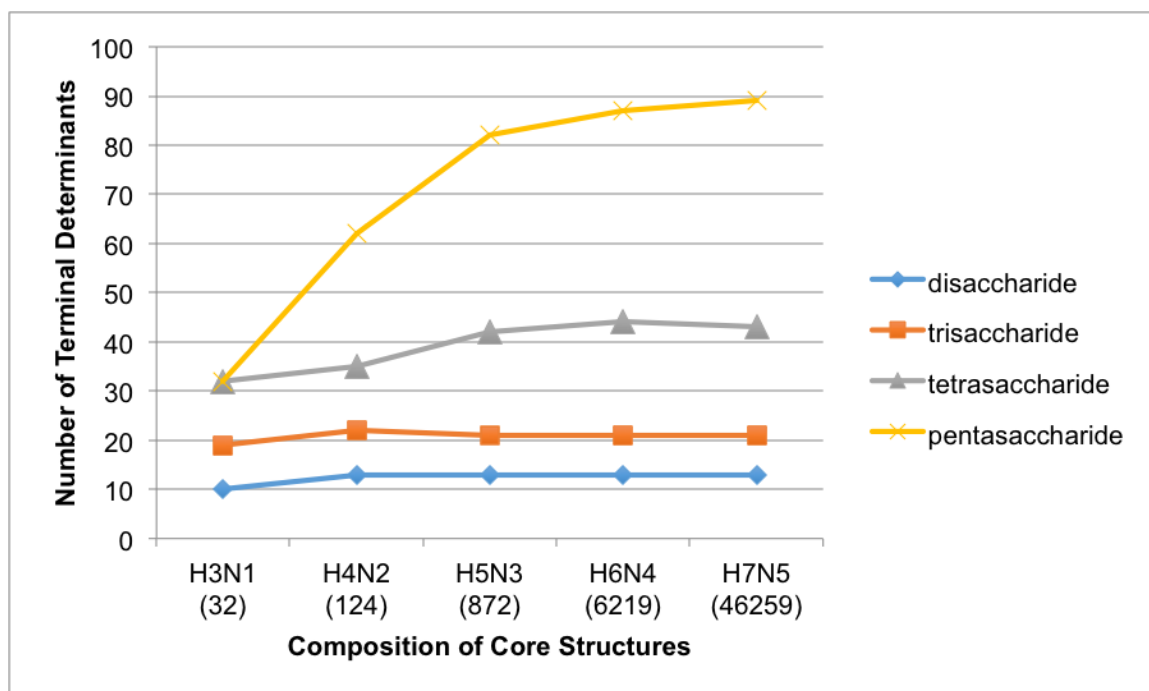


Figure 3.3: Number of terminal determinants in human milk glycan glycome as a function of increasing core structure size. The composition of the core structures is indicated by number of hexoses (H, galactose and reducing terminal glucose) and the number of N-acetylglucosamines (N) in each core structure. The number of unique glycan structures including 0 to 5 fucose residues and 0 to 2 sialic acid residues is indicated in parentheses under each core composition. The data show the number of di- (blue), tri- (green), tetra- (red), and pentasaccharide (violet) determinants found among the glycans comprising each core structure.

before and after treatment with the specific exoglycosidases, β 1-3-galactosidase, β 1-4-galactosidase, and α 1-2-fucosidase and endo- β 1-4-galactosidase are shown in Supplemental Table 1. The determinants (subset of a glycan structure containing di- to pentasaccharides) identified by defined GBP binding are shown in Supplemental Fig. 2.

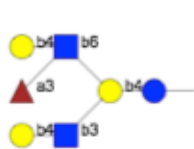
The application outputs a table with the predictions for each target and the associated metadata. To assist with the interpretation and derivation of the results, the application includes explanations for predicted structures and candidate structures that were ruled out as predicted structures due to filtering from the negative binders. Each predicted structure includes the metadata that was used to include it in the final result. For example, figure 3.4 shows the results of GlycomeSeq analysis of HMG-20 where the prediction is a single structure and three “Ruled out Candidates” are listed.

3.4 Discussion

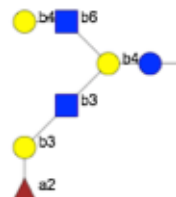
In the current studies we developed a novel computational approach to address the fundamental question of how large is the human milk metaglycome and can structural information and metadata be combined to facilitate high throughput sequencing of glycans. Predictions as to the number of different glycans that can exist in human milk have varied from a few hundred to many thousands. In our study we applied a novel approach to address that question as shown in Table 1 where we estimated that number of human milk glycans could be 53,000 possible structures, with the limitation of those to a core structure no larger than a dodecasaccharide. This estimate was based on current information available on the free-reducing human milk glycans and the enzymes involved in their synthesis. Obviously no single individual will produce all possible glycans because the structures synthesized will depend many factors including, the genotype, time after initiation of lactation, time of day, and

Sample ID	Composition	No. Virtual Candidates	Metadata	Ruled out Candidates	Prediction
HMG-20	H4N2F1S0	12	ECL Terminal Monofucosylated LacNAc Branched Lactose	Fa2Lb3Nb3(Lb4Nb6)Lb4G anti-type-1 Lb3(Fa4)Nb3(Lb4Nb6)Lb4G anti-Aea Lb4(Fa3)Nb3(Lb4Nb6)Lb4G anti-Aex	Ab4GNb3(Ab4(Fa3)GNb6)Ab4G

(a) Display of metadata and results from GlycomeSeq web application.



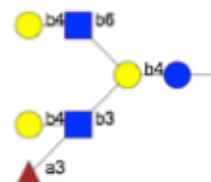
(b) The candidate structure for HMG-20 is positive for ECL, Terminal Monofucosylated LacNAc, and Branched Lactose



(c) The candidate structure for HMG-20 is positive for ECL, Terminal Monofucosylated LacNAc, and Branched Lactose but has been ruled out since it contains a determinant for a non-binding Glycan Binding Protein (anti-Type I).



(d) The candidate structure for HMG-20 is positive for ECL, Terminal Monofucosylated LacNAc, and Branched Lactose but has been ruled out since it contains a determinant for a non-binding Glycan Binding Protein (anti-Lea).



(e) The candidate structure for HMG-20 is positive for ECL, Terminal Monofucosylated LacNAc, and Branched Lactose but has been ruled out since it contains a determinant for a non-binding Glycan Binding Protein (anti-Lex).

Figure 3.4: GlycomeSeq display for HMG-20. A. The display shows the Sample ID (HMG-20), the composition and the number of isomers in the virtual glycome with that composition (Virtual Candidates). The positive metadata supporting the single prediction is shown and the structures of the Candidate glycans ruled out by non-binding lectins or antibodies are presented as “Ruled out Candidates”. Finally the single prediction is shown. All structures are presented in a linear code (ref). B. Clicking on the linear code under “Prediction” in 4A., displays the candidate structure and the positive metadata. If one clicks on the sample ID in 4B (HMG-20), the link to the structure in the Virtual Milk Glycome in the UniCarbKB database is displayed (not shown). C-E. are displayed when the linear code under Ruled out Candidates (3A.) are clicked on from top to bottom, respectively. Here the logic for ruling out the structure from the database is presented for each.

nutrition state of the donors. Nevertheless, it is clear that many thousands of different glycans are synthesized and secreted into the milk of all human mothers. Thus, modern computational approaches as developed here are essential to help identify and characterize the complex metaglycomes of human milk.

We hypothesized that a bioinformatics approach for structural analysis that combines the knowledge of a database of fully characterized glycan structures and experimental metadata from glycan microarrays and MS analysis would be able to automate the sequencing for the human milk meta-glycome. This approach is summarized in the algorithm GlycomeSeq (Fig. 2), and as shown in Supplemental Table 2, GlycomeSeq was able to identify all of the standards on the array that were found in the database of all human milk free glycans (virtual glycome). The algorithm correctly predicted no structures for “Agal LNT”, “LNFP IV” (H2), and “Ley-Lex” since they do not occur in human milk. GlycomeSeq was able to identify a single prediction in the unknown glycan targets in 20 out of 33 cases, and in all cases drastically reduced the number of candidate glycans from the large numbers of possible isomers in each composition. Structures of the glycan targets were confirmed or determined by independent structural methods and were in agreement with the predicted structures from GlycomeSeq.

Generally, where the algorithm produced multiple predictions for a target, there may not be enough meta-data available to eliminate candidate structures or the data indicates a binding motif should be present but is actually missing in a candidate structure. Such instances may occur due to weak binding, cross-reactivity of GBP for a target glycan, steric effects that prevent detection of a determinant that actually exists in the target, or insufficient data quality. For example, glycan HMG-76 with composition H6N4F2S0 has 291 possible structures in the virtual glycome. GlycomeSeq predicted 2 structures and all them contained determinants for the GBPs and the MSn reporter ions. Neither of the 2 predicted structures were ruled out during the

negative binder filtering-step. In spite of not being able to generate a single predicted structure, the algorithm was able to eliminate all but 2 candidate structures from the 291 possible isomers with this composition. Such information is invaluable to analysts using mass spectrometry to define this gly-can target and facilitates more targeted MS techniques to distinguish isomers.

The current state of the art approach for automated glycan sequencing is through techniques that automatically interpret MS data. However, these methods and tools are still evolving. Methods vary from (i) matching theoretical peak lists to the mass spectra [52], (ii) matching mass spectra to a database of experimentally determined spectra [53], (iii) de novo sequencing approaches that match mass spectra to theoretical peak lists from structures that constrained by biosynthetic pathways [34, 59, 38, 49], and (iv) semi-automated annotation to assist the manual interpretation of MS data [23]. All of these approaches are limited by their ability to perform complete structural characterizations that detect linkage position and anomeric configurations, meaning that the predicted glycan structures are ambiguous in certain aspects. Based on our current review of the literature, GlycomeSeq is the only automated high throughput sequencing method that can predict fully characterized glycan structures including topology, linkages, and anomeric configurations.

We also found that the number of terminal determinants that are found in human milk free glycans increases sub-linearly as a function of core structure size. This raises the possibility that the large numbers of isomeric structures represent a type of scaffold upon which specific determinants are created to provide necessary biological functions. From the analysis in Fig. 2 we observe that as the number of monosaccharides in the core structures of the glycans increases, the number of terminal determinants increases, which is consistent with the greater branching that can occur in the larger glycans. Interestingly, the number of tetra- and pentasaccharide determinants seems to reach a constant number at a core structure of an octasac-

charide, suggesting that free milk glycans up to octasaccharides may represent the biologically relevant set of glycans in human milk. These observations also suggest that the free glycans in human milk present a repertoire of structures that present biologically relevant determinants, and that individual structures are less important than the “bouquet” of determinants.

Adding orthogonal methods as metadata to GlycomeSeq enhances the predictive power of our method. The reporter ions from MSn analysis provide conclusive structural information for fragments in the unknown target structure. For example, HMG-76 has two reporter ions inferred from the MSn analysis, if we eliminate the fragments from our input then the algorithm predicts 47 structures from the binding data alone. Similarly, if we only use the reporter ion fragments from MS, GlycomeSeq predicts 10 structures. By combining the GBP binding data with the MS fragment data; the algorithm generates 2 predicted structures out of 291 possible structures.

3.5 Conclusion

We describe herein an approach to define a virtual meta-glycome and use it as a knowledge base to predict fully characterized glycan structures using data from MS and glycan microarray binding experiments. This approach to computational sequencing of the unknown glycans requires (a) determination of the glycan composition using MALDI-TOF analysis, (b) interrogation of the glycans with lectins and antibodies that bind known determinants, (c) determination of the set of predicted structures based on automated meta-analysis of the experimental data from the virtual glycome database given the constraints of the rules for the biosynthetic pathway of the glycans. While several methods have been aimed at glycan annotation of mass spectrometric analysis, to our knowledge no other method has been developed that attempts to solve the glycan structure to the level of GlycomeSeq. This approach has the potential to be a significant breakthrough in glycomics analysis that has thus far

been hindered by the complexity and ambiguity of MS analysis. We seek to make improvements and have planned future work for our approach including (i) automate the analysis of the MS spectra to identify the MS reporter ion determinants, (ii) develop a library of XPath queries for the determinants so the user does not have to manually specify them in the spreadsheet, (iii) include exo- and endoglycosidases to yield finer specificity, and (iv) apply this method on HMGs with Sialic Acid to further validate this approach before moving onto a much larger metaglycome.

Chapter 4

A Web-Based Bioinformatics Platform for Glycan Microarray Analysis

4.1 Introduction

Recognizing the need to take an integrated approach to advance glycan structure-function relationships, several international collaborative efforts, including the Consortium for Functional Glycomics (CFG), EuroCarb, the Japanese Consortium for Glycomics, and many other resources have been established to develop novel resources and technologies for glycomics [80]. The bioinformatics resource accessible at the CFG Functional Glycomics Gateway (<http://www.functionalglycomics.org>) integrates data from studies on glyco-gene microarrays, mass spectrometry mass profiles of cells and tissues, transgenic mice with knockouts of glycan biosynthetic enzymes, and glycan microarrays interrogated with GBPs. All of these data sets provide a systems framework for glycomics to integrate information from the molecular to organism level. One of the most widely used resources, the publicly available defined glycan microarray now has over 600 glycan targets and has generated over 5,500 primary screen analyses from over 1,000 unique proteins during the past ten years. The printed glycan microarray is available to investigators worldwide for exploring GBP specificities as part of its mission to define the paradigms by which protein-carbohydrate interactions mediate cell communication.

While the CFG is the largest resource of glycan microarray data that we are aware of, determination of GBP specificity is a subjective and manual process re-

quiring expert analysis. In the microarray format, GBP specificity is predicted by inspecting the binding intensity of GBPs to the different glycans on the array and further identifying a glycan motif. The manual and subjective inspection of the binding data motivated us to develop a computational approach to automate the detection of GBP binding specificity to discover glycan motifs as discussed in [chapter 2](#).

The CFG also has glycan profiling resources that used MALDI mass spectrometry and other analyses to identify and characterize the glycans from glycoconjugates in human and mouse tissues and cells. However, similar to determining GBP specificity, the annotation of glycan structures is a low-throughput process and is limited by its ability to characterize the fine structure of the glycan. We proposed a computational method to fully sequence glycans using MAGS and described this approach in [chapter 3](#). While other computational methods exist for glycan motif mining [115, 3] and annotating glycan structures based on mass profiles [38, 23], the lack of user-friendly tools accessible for the average glycobiochemist researcher still persists. The reliance of computational tools can be difficult for life science researchers to install, use, and interpret results. However, one needs not look further than genomics and cancer genomics to take notice of the impact of bioinformatics tools and platforms [69, 43] that have become increasingly more accessible to life science researchers. Motivated by the complexity of glycomics analysis and the tremendous success of genomic bioinformatics platforms, we developed the web-based bioinformatics platform, termed GlycoPattern, to support reproducible and transparent research with user friendly tools. We describe herein the latest version of GlycoPattern that includes support for a plugin framework for motif mining algorithms and supports mammalian glycan microarrays beyond the CFG.

4.2 Methods

GlycoPattern is an open-source, publicly available web-based resource that allows investigators to store and analyze glycan microarray data. It features the ability to determine the specificities of GBPs, interactive tools to view and search experimental data, and a heatmap tool to compare multiple glycan microarray experiments to determine the fine specificity of GBPs. The software tools used to develop GlycoPattern include Python (v2.7) and the Pylons web framework (v1.0) for the server-side code and HTML, Javascript, and CSS for the front-end code. GlycoPattern uses a relational database (MySQL v5.6) to store the glycan structures, array versions, and experiments. The project is open source and the source code is available at <http://github.org/sagravat/glycopattern>. Various features will be described in the following sections.

Glycan Notation Many notations representing textual glycan nomenclature are available, including International Union of Pure and Applied Chemistry (IUPAC) [70], Linear Notation for Unique description of Carbohydrate Structures (LINUCS; [18], Kyoto Encyclopedia of Genes and Genomes (KEGG) Chemical Function [46], LinearCode [31], GLYDE-II [74] and GlycoCT [47]. We chose to use the modified IUPAC condensed nomenclature from the CFG because of the wide adoption of IUPAC in the Glycobiology community and the inclusion of the anomeric carbon. For structure searches, GlycoPattern can accept and display the modified IUPAC nomenclature for oligosaccharides.

Symbolic Representation of Glycans As described in the GlycanBuilder tool [22], the symbolic representation of glycans using a “cartoon” format consists of a series of geometric symbols that represent monosaccharide units connected by lines to indicate glycosidic linkages. The GlycanBuilder is an applet for building and displaying glycan structures, however, the lack of an embeddable HTML5 compliant

solution prompted us to develop a Javascript and HTML5-based solution using HTML Canvas. The advantage of the HTML5 approach is the speed of rendering the symbols without requiring the need to store an image file on disk for each glycan structure or substructure. Our Javascript code takes an IUPAC code as input and generates an HTML canvas element containing the symbols.

Array Versions The previous version of GlycoPattern was limited to the CFG glycan microarrays with predefined glycans. Our latest version has been extended to support custom developed mammalian glycan microarrays. Users can use one of the predefined microarrays or create a custom glycan microarray by entering the glycan identifier (short name or numerical identifier) and the abbreviated IUPAC encoding of the glycan structure for each structure. The database contains a table to store the glycan structures and a separate table for the glycan array version and a many-to-many linking table for glycans to arrays.

Experiments Experimental data in GlycoPattern currently requires the user to store the summary target RFU value for each glycan on the microarray. Experiments are stored in a table that contains the name of the experiment, the concentration of the GBP, the microarray ID, and the user ID. The results of the experiment are stored in a one-to-many table of experiment to results that contains the experiment ID, the glycan ID, and the RFU value. To add or upload experimental data, users can copy and paste the average RFU values for the glycan targets from an experiment into GlycoPattern for the selected array version.

Data Normalization and Z-score After experimental data is entered, the data is normalized between 0.0 and 1.0 using a square root transformation. Since glycan microarray data frequently has RFU values that are not positive, all values less than or equal to 0 are set to 0 to simplify the transformation. The normalized values enable users to compare experiments for the same GBP at different concentrations and also compare the same glycan across different experiments. It is important to use

discretion when comparing the normalized values across different arrays as the values are relative to the glycans that bound on the microarray, thus are not an absolute value. The cutoff for high intensity binders is determined by a z-score > 1.96 on the normalized data. The cutoff value is stored in the experiment table to be used for visualization purposes when viewing the experimental data on a chart.

Motif Discovery A glycan array is interrogated with a GBP to determine the relative binding strengths of all the glycans on the array [92]. The previous version of GlycoPattern embedded the GlycanMotifMiner (GLYMMR) algorithm [25] to discover glycan motifs. The algorithm finds frequently occurring patterns (minimum of a disaccharide motif) in binding glycans using a threshold for the minimum number of glycans that contain the motif along with a threshold for the maximum number of non-binding glycans that contain the motif. As we develop new motif mining methods we consider making GlycoPattern flexible enough to “plug-in” other algorithms using RESTful Services. These algorithms could be implemented in any language provided the algorithm has an endpoint that is web-accessible and follows the request and response entity specifications. Thus, the choice of programming languages such as R, Java, Scala, in addition to Python, need not be a barrier to integrating with GlycoPattern. We use JavaScript Object Notation (JSON) as the input and output format for the REST service. The input object has a “binder” and “nonbinder” element. Each element contains a list of “GlycanArrayData” objects that include the glycan id, linear code, and the RFU value. The output object is a list of objects that includes the motif, the reference of high binder “GlycanArrayData” objects that contain the motif, the reference of non binder “GlycanAraryData” objects that contain the motif, and an optional p-value and q-value. The reference list of objects returned for high binders and non-binders is made accessible in the user interface to allow the user to view those particular glycans that include the motif to allow manual analysis of the glycans.

4.3 Results

The web-based visualization tools developed for GlycoPattern enable users to observe patterns and anomalies in the data using interactive web components. The visualization features of GlycoPattern allow users to view experimental data with an interactive chart, compare multiple experiments using a heatmap, and view glycan motifs for an experiment. GlycoPattern also supports searching glycan substructures within an experiment to provide users the capability to perform custom queries or filtering on the data based on a substructure. We describe the features in the following sections.

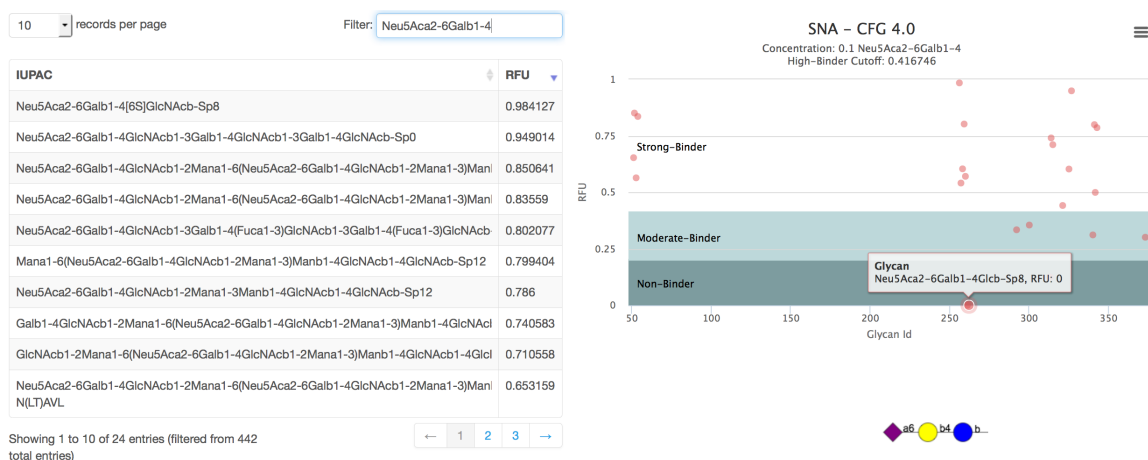
Experiment Chart The interactive experiment chart figure 4.1 allows users to view a scatter plot of the data and click on each point to view the cartoon of the structure. The table on the left is presorted by the normalized RFU and allows filtering based on the IUPAC code. The chart includes a shaded area to differentiate between strong binders and non-binders. Users can also download an image of the chart and save it to their computer.

GBP-glycan heatmap The GBP-glycan heatmap figure 4.2 feature allows the user to select two or more experiments (including experiments from different Array versions) and perform a side-by-side comparison of the glycans binding intensity for any number of GBPs. GlycoPattern allows the user to hover over a cell in the Heatmap and display the glycan structure in symbolic representation. This feature is useful for comparing the binding of different but related GBPs to explore the fine specificity of different GBPs.

Glycan search GlycoPattern allows users to search for glycan substructures within an experiment. The user interface requires the use of the modified IUPAC condensed nomenclature. The search supports hierarchical queries that represent the branching in certain glycan sequences where branching is noted with an opening and



(a) full chart




(b) filtered chart

Figure 4.1: The full chart in (a) allows the user to see the data in a sorted table and in a scatterplot. The user can click on a point in the scatter plot and the plot will render the glycan structure below the figure as shown. The user can click on the three striped lines above the chart to download the figure. The filtered chart (b) shows the table filtered by all structures that contain Neu5Aca2-6Galb1-4. The chart on the right also reflects the structures that contain the same filtered data. One of the non-binders has been selected and shown.

Column Header Key: 0 - UEA 1 - v4.0 1 - AAL - v4.0

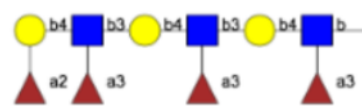
Glycan Id	0	1
73Sp8		96
70Sp8	15	92
67Sp0	70	85
11Sp8		80
62Sp0		80
66Sp0	26	78
64Sp8	41	74
260Sp0	95	66
84Sp8	58	65
68Sp0	65	62
80Sp18		62
107Sp8	74	60
213Sp8	81	59
32Sp8		54
61Sp0		54
69Sp8	13	54
252Sp0		52
67Sp8	79	51
115Sp8		51
211Sp8		51
212Sp0	56	48
11Sp8		47



(a) The 3rd glycan from the top is 67Sp0, which is H-type 2 structure shown in symbols on the right. These data indicate that AAL and UEA-1 both bind this structure.

Column Header Key: 0 - UEA 1 - v4.0 1 - AAL - v4.0

Glycan Id	0	1
73Sp8		96
70Sp8	15	92
67Sp0	70	85
11Sp8		80
62Sp0		80
66Sp0	26	78
64Sp8	41	74
260Sp0	95	66
84Sp8	58	65
68Sp0	65	62
80Sp18		62
107Sp8	74	60
213Sp8	81	59
32Sp8		54
61Sp0		54
69Sp8	13	54



(b) The 5th glycan from the top is 62Sp0, which is a fucosylated polylactosamine structure that is bound strongly by AAL (80%) but unbound by UEA-I. This is a useful feature for comparing the fine specificities of glycan binding proteins.

Figure 4.2

closing parenthesis around a sequence. The search function parses the modified IUPAC condensed nomenclature and converts it into Javascript Object Notation (JSON) format. When searching for a branched input sequence, the search function finds any match within a glycan structure for the input sequence rather than an exact match. This is useful when dealing with sequences that have more than two branch points from a single monosaccharide. The Search feature allows the user to select and enter a specific motif or determinant on the array and the program will display all of the defined glycans on the array from a selected experiment or data set that contain the entered determinant in order of their normalized RFU value. An example of using the Glycan Search feature is illustrated in figure 4.3

Instructions: Enter the IUPAC code in the text box and select an experiment to search against.

<input type="text" value="Neu5Aca2-6Galb1-4Glc"/>	<input type="text" value="SNA - 5.0"/>	<input type="button" value="Submit"/>
2 RESULTS FOUND		
[271Sp8] Neu5Aca2-6Galb1-4Glc-Sp8	[271Sp0] Neu5Aca2-6Galb1-4Glc-Sp0	
0.0890214	0	



Figure 4.3: Enter any structure or motif using IUPAC nomenclature; i.e., we might want to see how 6'-sialyl-lactose is bound by SNA. To do this enter Neu5Aca2-6Galb1-4Glc and click Submit. Only two structures are found on the array have this structure and they are 6'-Sialylactose with two different linkers and neither are bound efficiently on the array. SNA has a low affinity for this determinant when it has Glucose on the reducing end.

4.4 Discussion

GlycoPattern is one of the few publicly available informatics resources available for mining Glycan Array data. At the time of writing, there are approximately 130 international registered users with over 1200 experiments total. It was originally developed as a resource for the CFG, but has been expanded beyond the CFG to support custom mammalian glycan microarrays. GlycoPattern also supports pluggable algorithms developed in any programming language. This flexibility will facilitate integration with previous motif mining methods or new methods that may not have been developed in the same programming language as GlycoPattern. The interactive features of GlycanPattern that allow for comparison of multiple experiments enable exploring the fine specificity of closely related but different GBPs.

In our planned future work, GlycoPattern will be adapted to support uploading the raw text output files of certain microarray scanners. The raw output data contains enriched data including glycan target replicate RFU values and background RFU values. Efforts such as the joint international consortium for the Minimum Information Required for a Glycomics Experiment [116] are actively working on developing standards to represent the minimum information required for a glycan microarray experiment similar to the efforts of the Minimum Information Required for a Microarray Experiment [97, 81]. Once standards are adopted, storage, sharing, querying and analysis of glycan microarray data will be enriched and lead to more maturity in the development of bioinformatics tools for the glycobiology community.

Chapter 5

Conclusion and Future Work

The need for informatics resources and tools for glycosciences has become increasingly evident to help experimentalists understand structure-function relationships in glycans, perform glycan sequencing/structural characterizations, and making glycomics information accessible to the scientific community. The complexity of glycan structures requires the development of unique bioinformatics resources and tools beyond proteomics and genomics. In this thesis, we have attempted to demonstrate approaches aimed at discovering the various structure and function relationships of glycans and GBPs using glycan microarrays and computational methods. In [chapter 2](#), we found that computational methods can serve as a hypothesis generating platform to determine the specificities of GBPs and glycans. In [chapter 3](#) we demonstrated a computational approach to sequencing glycan structures to a level of detail beyond any other automated approach that we are aware of. Finally, in [chapter 4](#), we demonstrated an informatics resource to allow investigators to store, analyze, and visualize glycan microarray experimental data. We extended GlycoPattern beyond the CFG to support glycan microarrays for any mammalian based glycans.

5.1 Future Work

Building on the informatics tools and methods reviewed and presented in this thesis, we plan to extend our work by to develop informatics for cross-platform gly-

can array data integration and mining. This will include defining a format that is amenable to digital conversion of public glycan array data and for machine access of the data so as to exchange datasets between different databases. We will also develop tools to convert the CFG glycan array repository and to import glycan array datasets generated on other platforms from literature into this format. Subsequent glycan array datasets on different array formats generated by various groups could then automatically be imported into this platform for storage, analysis, and sharing.

A variety of software tools for glycan microarray platforms will be developed. We will also build on our previous efforts to develop algorithms to mine binding motifs from datasets generated across multiple glycan microarray platforms. The method for sequencing glycans will be applied to larger and more diverse glycomes starting with the sialyated human milk glycans and eventually applied to the N-Glycans.

Accomplishing these aims would build on current efforts and lay the foundation for building larger programs to support the development of robust bioinformatics in accordance to the comprehensive roadmap laid out by the National Academy of Sciences to advance glycomics.

Bibliography

- [1] Sanjay B Agravat, Joel H Saltz, Richard D Cummings, and David F Smith. Glycopattern: a web platform for glycan array mining. Bioinformatics, 30(23):3417–3418, 2014. 18
- [2] Y. Akune, M. Hosoda, S. Kaiya, D. Shinmachi, and K. F. Aoki-Kinoshita. The rings resource for glycome informatics analysis and data mining on the web. OMICS, 14(4):475–86, 2010. 13
- [3] K. F. Aoki-Kinoshita. Mining frequent subtrees in glycan data using the rings glycan miner tool. Methods Mol Biol, 939:87–95, 2013. 15, 74
- [4] Kiyoko F. Aoki-Kinoshita. Glycome Informatics: Methods and Applications. CRC Press, 2009. 7
- [5] N. V. Artemenko, A. G. McDonald, G. P. Davey, and P. M. Rudd. Databases and tools in glycobiology. Methods Mol Biol, 899:325–50, 2012.
- [6] D. J. Ashline, A. J. Hanneman, H. Zhang, and V. N. Reinhold. Structural documentation of glycan epitopes: sequential mass spectrometry and spectral matching. J Am Soc Mass Spectrom, 25(3):444–53, 2014. 55
- [7] D. J. Ashline, Y. Yu, Y. Lasanajak, X. Song, L. Hu, S. Ramani, V. Prasad, M. K. Estes, R. D. Cummings, D. F. Smith, and V. N. Reinhold. Structural characterization by multistage mass spectrometry (msn) of human milk glycans recognized by human rotaviruses. Mol Cell Proteomics, 13(11):2961–74, 2014. 56, 58
- [8] J. Axford. The impact of glycobiology on medicine. Trends Immunol, 22(5):237–9, 2001. 1
- [9] J. U. Baenziger and D. Fiete. Structural determinants of concanavalin a specificity for oligosaccharides. J Biol Chem, 254(7):2400–7, 1979. 38, 40
- [10] S. E. Baldus, J. Thiele, Y. O. Park, F. G. Hanisch, J. Bara, and R. Fischer. Characterization of the binding specificity of anguilla anguilla agglutinin (aaa) in comparison to ulex europaeus agglutinin i (uea-i). Glycoconj J, 13(4):585–90, 1996. 40

- [11] C. R. Bertozzi and D. Rabuka. Structural Basis of Glycan Diversity. Cold Spring Harbor (NY), 2nd edition, 2009. 54
- [12] J. Bienenstock, R. H. Buck, H. Linke, P. Forsythe, A. M. Stanisz, and W. A. Kunze. Fucosylated but not sialylated milk oligosaccharides diminish colon motor contractions. Plos One, 8(10), 2013. 56
- [13] M. F. Bierhuizen, M. G. Mattei, and M. Fukuda. Expression of the developmental i antigen by a cloned human cDNA encoding a member of a beta-1,6-n-acetylglucosaminyltransferase gene family. Genes Dev, 7(3):468–78, 1993.
- [14] G. W. Bird. Anti-t in peanuts. Vox sanguinis, 9:748–9, 1964. 35
- [15] O. Blixt, S. Head, T. Mondala, C. Scanlan, M. E. Huflejt, R. Alvarez, M. C. Bryan, F. Fazio, D. Calarese, J. Stevens, N. Razi, D. J. Stevens, J. J. Skehel, I. van Die, D. R. Burton, I. A. Wilson, R. Cummings, N. Bovin, C. H. Wong, and J. C. Paulson. Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. Proc Natl Acad Sci U S A, 101(49):17033–8, 2004. 13, 19, 21
- [16] O. Blixt and N. Razi. Chemoenzymatic synthesis of glycan libraries. Methods Enzymol, 415:137–53, 2006. 2
- [17] L. Bode. Human milk oligosaccharides: Every baby needs a sugar mama. Glycobiology, 22(9):1147–1162, 2012. 56, 57
- [18] A. Bohne-Lang, E. Lang, T. Forster, and C. W. von der Lieth. Linucs: linear notation for unique description of carbohydrate sequences. Carbohydr Res, 336(1):1–11, 2001. 75
- [19] M. P. Campbell, R. Ranzinger, T. Lutteke, J. Mariethoz, C. A. Hayes, J. Zhang, Y. Akune, K. F. Aoki-Kinoshita, D. Damerell, G. Carta, W. S. York, S. M. Haslam, H. Narimatsu, P. M. Rudd, N. G. Karlsson, N. H. Packer, and F. Lisacek. Toolboxes for a standardised and systematic study of glycans. BMC Bioinformatics, 15 Suppl 1:S9, 2014. 63
- [20] S. Carlsson, C. T. Oberg, M. C. Carlsson, A. Sundin, U. J. Nilsson, D. Smith, R. D. Cummings, J. Almkvist, A. Karlsson, and H. Leffler. Affinity of galectin-8 and its carbohydrate recognition domains for ligands in solution and at the cell surface. Glycobiology, 17(6):663–76, 2007. 44
- [21] E. Castanys-Munoz, M. J. Martin, and P. A. Prieto. 2'-fucosyllactose: an abundant, genetically determined soluble glycan present in human milk. Nutr Rev, 71(12):773–89, 2013.
- [22] Alessio Ceroni, Anne Dell, and Stuart M Haslam. Source code for biology and medicine. Source code for biology and medicine, 2:3, 2007. 75

- [23] Alessio Ceroni, Kai Maass, Hildegard Geyer, Rudolf Geyer, Anne Dell, and Stuart M Haslam. Glycoworkbench: a tool for the computer-assisted annotation of mass spectra of glycans. Journal of proteome research, 7(4):1650–1659, 2008. 70, 74
- [24] Yun Chi, Richard R Muntz, Siegfried Nijssen, and Joost N Kok. Frequent subtree mining-an overview. Fundamenta Informaticae, 66(1-2):161–198, 2004. 20
- [25] S. R. Cholleti, S. Agravat, T. Morris, J. H. Saltz, X. Z. Song, R. D. Cummings, and D. F. Smith. Automated motif discovery from glycan array data. Omicron-a Journal of Integrative Biology, 16(10):497–512, 2012. 18, 77
- [26] R. D. Cummings. Use of lectins in analysis of glycoconjugates. Methods Enzymol, 230:66–86, 1994. 38
- [27] R. D. Cummings. The repertoire of glycan determinants in the human glycome. Mol Biosyst, 5(10):1087–104, 2009. 9, 19, 54, 63
- [28] T. De Vries, M. P. Palcic, P. S. Schoenmakers, D. H. Van Den Eijnden, and D. H. Joziase. Acceptor specificity of gdp-fuc:gal beta 1->4glcnac-r alpha 3-fucosyltransferase vi (fuct vi) expressed in insect cells as soluble, secreted enzyme. Glycobiology, 7(7):921–7, 1997.
- [29] H. Debray, D. Decout, G. Strecker, G. Spik, and J. Montreuil. Specificity of twelve lectins towards oligosaccharides and glycopeptides related to n-glycosylproteins. Eur J Biochem, 117(1):41–55, 1981. 40
- [30] G. Duska-McEwen, A.P. Senft, T.L. Ruetschilling, E.G. Barrett, and R.H. Buck. Oligosaccharides enhance innate immunity to respiratory syncytial virus and influenza in vitro. Food and Nutrition Sciences, 5:1387–1398, 2014. 56
- [31] B. EHUD, N. Yael, A. Yaniv, H. Asaf, I. Ori, N. Dotan, and D. Avinoam. A novel linear code nomenclature for complex carbohydrates. Trends Glycoscience Glycotechnology, 14(77):127–137 0915–7352, 2002. 60, 75
- [32] T. Feizi and W. Chai. Oligosaccharide microarrays to decipher the glyco code. Nature reviews. Molecular cell biology, 5(7):582–8, 2004. 19
- [33] S. Fukui, T. Feizi, C. Galustian, A. M. Lawson, and W. Chai. Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. Nat Biotechnol, 20(10):1011–7, 2002. 19
- [34] Sara P Gaucher, Jeff Morrow, and Julie A Leary. Stat: a saccharide topology analysis tool used in combination with tandem mass spectrometry. Analytical chemistry, 72(11):2331–2336, 2000. 55, 70
- [35] Transforming Glycoscience. A roadmap for the future, 2012. 3

- [36] Jeremy Goecks, Anton Nekrutenko, James Taylor, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol, 11(8):R86, 2010.
- [37] K. C. Goehring, A. D. Kennedy, P. A. Prieto, and R. H. Buck. Direct evidence for the presence of human milk oligosaccharides in the circulation of breastfed infants. Plos One, 9(7), 2014. 57
- [38] David Goldberg, Mark Sutton-Smith, James Paulson, and Anne Dell. Automatic annotation of matrix-assisted laser desorption/ionization n-glycan spectra. PROTEOMICS, 5(4):865–875, 2005. 16, 55, 70, 74
- [39] I. J. Goldstein. Studies on the combining sites of concanavalin a. Advances in experimental medicine and biology, 55:35–53, 1975. 38, 40
- [40] I. J. Goldstein, C. E. Hollerman, and E. E. Smith. Protein-carbohydrate interaction. ii. inhibition studies on the interaction of concanavalin a with polysaccharides. Biochemistry, 4:876–83, 1965. 38, 40, 51
- [41] E. F. Grollman and V. Ginsburg. Correlation between secretor status and the occurrence of 2'-fucosyllactose in human milk. Biochem Biophys Res Commun, 28(1):50–3, 1967.
- [42] E. F. Grollman, A. Kobata, and V. Ginsburg. An enzymatic basis for lewis blood types in man. The Journal of clinical investigation, 48(8):1489–94, 1969.
- [43] David A Gutman, Jake Cobb, Dhananjaya Somanna, Yuna Park, Fusheng Wang, Tahsin Kurc, Joel H Saltz, Daniel J Brat, Lee AD Cooper, and Jun Kong. Cancer digital slide archive: an informatics resource to support integrated in silico analysis of tcga pathology data. Journal of the American Medical Informatics Association, 20(6):1091–1098, 2013. 74
- [44] S. Hammarstrom and E. A. Kabat. Studies on specificity and binding properties of the blood group a reactive hemagglutinin from helix pomatia. Biochemistry, 10(9):1684–92, 1971. 32
- [45] K. Hashimoto, I. Takigawa, M. Shiga, M. Kanehisa, and H. Mamitsuka. Mining significant tree patterns in carbohydrate sugar chains. Bioinformatics, 24(16):i167–73, 2008. 20, 47
- [46] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. J Am Chem Soc, 125(39):11853–65, 2003. 75
- [47] S. Herget, R. Ranzinger, K. Maass, and C. W. Lieth. Glycoct-a unifying sequence format for carbohydrates. Carbohydr Res, 343(12):2162–71, 2008. 75

- [48] H. D. Holscher, S. R. Davis, and K. A. Tappenden. Human milk oligosaccharides influence maturation of human intestinal caco-2bbe and ht-29 cell lines. J Nutr, 144(5):586–91, 2014. 57
- [49] Han Hu, Yu Huang, Yang Mao, Xiang Yu, Yongmei Xu, Jian Liu, Chengli Zong, Geert-Jan Boons, Cheng Lin, Yu Xia, et al. A computational framework for heparan sulfate sequencing using high-resolution tandem mass spectra. Molecular & Cellular Proteomics, 13(9):2490–2502, 2014. 70
- [50] E. Jantscher-Krenn, T. Lauwaet, L. A. Bliss, S. L. Reed, F. D. Gillin, and L. Bode. Human milk oligosaccharides reduce entamoeba histolytica attachment and cytotoxicity in vitro. British Journal of Nutrition, 108(10):1839–1846, 2012. 56
- [51] P. H. Johnson and W. M. Watkins. Purification of the lewis blood-group gene associated alpha-3/4-fucosyltransferase from human milk: an enzyme transferring fucose primarily to type 1 and lactose-based oligosaccharide chains. Glycoconj J, 9(5):241–9, 1992.
- [52] Hiren J Joshi, Mathew J Harrison, Benjamin L Schulz, Catherine A Cooper, Nicolle H Packer, and Niclas G Karlsson. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. Proteomics, 4(6):1650–1664, 2004. 70
- [53] Akihiko Kameyama, Norihiro Kikuchi, Shuuichi Nakaya, Hiromi Ito, Takashi Sato, Toshihide Shikanai, Yoriko Takahashi, Katsutoshi Takahashi, and Hisashi Narimatsu. A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. Analytical chemistry, 77(15):4719–4725, 2005. 70
- [54] R. J. Kelly, S. Rouquier, D. Giorgi, G. G. Lennon, and J. B. Lowe. Sequence and expression of a candidate for the human secretor blood group alpha(1,2)fucosyltransferase gene (fut2). homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. J Biol Chem, 270(9):4640–9, 1995.
- [55] N. Kikuchi, A. Kameyama, S. Nakaya, H. Ito, T. Sato, T. Shikanai, Y. Takahashi, and H. Narimatsu. The carbohydrate sequence markup language (cabosml): an xml description of carbohydrate structures. Bioinformatics, 21(8):1717–8, 2005.
- [56] K. L. Koszdin and B. R. Bowen. The cloning and expression of a human alpha-1,3 fucosyltransferase capable of forming the e-selectin ligand. Biochem Biophys Res Commun, 187(1):152–7, 1992.
- [57] L. Kuhn, G. M. Aldrovandi, M. Sinkala, C. Kankasa, K. Semrau, M. Mwiya, P. Kasonde, N. Scott, C. Vwalika, J. Walter, M. Bulterys, W. Y. Tsai, and D. M.

- Thea. Effects of early, abrupt weaning on hiv-free survival of children in zambia (vol 359, pg 130, 2008). New England Journal of Medicine, 359(17):1859–1859, 2008. 56, 57
- [58] J. F. Kukowska-Latallo, R. D. Larsen, R. P. Nair, and J. B. Lowe. A cloned human cDNA determines expression of a mouse stage-specific embryonic antigen and the lewis blood group alpha(1,3/1,4)fucosyltransferase. Genes Dev, 4(8):1288–303, 1990.
- [59] Anthony J Lapadula, Philip J Hatcher, Andy J Hanneman, David J Ashline, Hailong Zhang, and Vernon N Reinhold. Congruent strategies for carbohydrate sequencing. 3. oscar: An algorithm for assigning oligosaccharide topology from ms n data. Analytical chemistry, 77(19):6271–6279, 2005. 70
- [60] R. D. Larsen, L. K. Ernst, R. P. Nair, and J. B. Lowe. Molecular cloning, sequence, and expression of a human gdp-l-fucose:beta-d-galactoside 2-alpha-l-fucosyltransferase cDNA that can form the h blood group antigen. Proc Natl Acad Sci U S A, 87(17):6674–8, 1990.
- [61] J. Loscalzo, I. Kohane, and A. L. Barabasi. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol, 3:124, 2007. 1
- [62] R. Lotan, E. Skutelsky, D. Danon, and N. Sharon. The purification, composition, and specificity of the anti-t lectin from peanut (*arachis hypogaea*). J Biol Chem, 250(21):8518–23, 1975. 35, 51
- [63] J. B. Lowe. The blood group-specific human glycosyltransferases. Baillieres Clin Haematol, 6(2):465–92, 1993.
- [64] J. Maksimovic, J. A. Sharp, K. R. Nicholas, B. G. Cocks, and K. Savin. Conservation of the st6gal i gene and its expression in the mammary gland. Glycobiology, 21(4):467–81, 2011.
- [65] I. Matsumoto and T. Osawa. Purification and characterization of an anti-h(o) phytohemagglutinin of *ulex europaeus*. Biochimica et biophysica acta, 194(1):180–9, 1969. 24, 40, 42, 52
- [66] P. Mattila, H. Salminen, L. Hirvas, J. Niittymaki, H. Salo, R. Niemela, M. Fukuda, O. Renkonen, and R. Renkonen. The centrally acting beta1,6n-acetylglucosaminyltransferase (glcnac to gal). functional expression, purification, and acceptor specificity of a human enzyme involved in midchain branching of linear poly-n-acetyllactosamines. J Biol Chem, 273(42):27633–9, 1998.
- [67] K. A. Maupin, D. Liden, and B. B. Haab. The fine-specificity of mannose-binding and galactose-binding lectins revealed using outlier-motif analysis of glycan array data. Glycobiology, 2011. 14

- [68] K. A. Maupin, D. Liden, and B. B. Haab. The fine specificity of mannose-binding and galactose-binding lectins revealed using outlier motif analysis of glycan array data. *Glycobiology*, 22(1):160–9, 2012. 20, 24, 47
- [69] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogiannakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008. 74
- [70] A. D. McNaught. Nomenclature of carbohydrates (recommendations 1996). *Adv Carbohydr Chem Biochem*, 52:43–177, 1997. 75
- [71] R. K. Merkle and R. D. Cummings. Lectin affinity chromatography of glycopeptides. *Methods Enzymol*, 138:232–59, 1987. 38
- [72] B. Mulloy, G. W. Hart, and P. Stanley. *Structural Analysis of Glycans*. Cold Spring Harbor (NY), 2nd edition, 2009. 54
- [73] J. Nakayama, J. C. Yeh, A. K. Misra, S. Ito, T. Katsuyama, and M. Fukuda. Expression cloning of a human alpha1, 4-n-acetylglucosaminyltransferase that forms glcnacalpha1->4galbeta->r, a glycan specifically expressed in the gastric gland mucous cell-type mucin. *Proc Natl Acad Sci U S A*, 96(16):8991–6, 1999. 32, 50
- [74] N. H. Packer, C. W. von der Lieth, K. F. Aoki-Kinoshita, C. B. Lebrilla, J. C. Paulson, R. Raman, P. Rudd, R. Sasisekharan, N. Taniguchi, and W. S. York. Frontiers in glycomics: bioinformatics and biomarkers in disease. an nih white paper prepared from discussions by the focus groups at a workshop on the nih campus, bethesda md (september 11-13, 2006). *Proteomics*, 8(1):8–20, 2008. 75
- [75] J. C. Paulson and K. J. Colley. Glycosyltransferases. structure, localization, and control of cell type-specific glycosylation. *J Biol Chem*, 264(30):17615–8, 1989.
- [76] J. Petryniak and I. J. Goldstein. Immunochemical studies on the interaction between synthetic glycoconjugates and alpha-l-fucosyl binding lectins. *Biochemistry*, 25(10):2829–38, 1986. 40, 44
- [77] A. Porter, T. Yue, L. Heeringa, S. Day, E. Suh, and B. B. Haab. A motif-based analysis of glycan array data to determine the specificities of glycan-binding proteins. *Glycobiology*, 20(3):369–80, 2010. 14, 20, 24, 47
- [78] A. K. Powell, Z. L. Zhi, and J. E. Turnbull. Saccharide microarrays for high-throughput interrogation of glycan-protein binding interactions. *Methods in molecular biology*, 534:313–29, 2009. 19

- [79] J. Quackenbush. Open-source software accelerates bioinformatics. Genome Biol, 4(9):336, 2003. 2
- [80] R. Raman, M. Venkataraman, S. Ramakrishnan, W. Lang, S. Raguram, and R. Sasisekharan. Advancing glycomics: implementation strategies at the consortium for functional glycomics. Glycobiology, 16(5):82R–90R, 2006. 73
- [81] T. F. Rayner, P. Rocca-Serra, P. T. Spellman, H. C. Causton, A. Farne, E. Holloway, R. A. Irizarry, J. Liu, D. S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, Jr. Stoeckert, C. J., J. White, P. L. Whetzel, F. Wymore, H. Parkinson, U. Sarkans, C. A. Ball, and A. Brazma. A simple spreadsheet-based, miame-supportive format for microarray data: Mage-tab. BMC Bioinformatics, 7:489, 2006. 82
- [82] O. Renkonen. Enzymatic in vitro synthesis of i-branches of mammalian polylactosamines: generation of scaffolds for multiple selectin-binding saccharide determinants. Cell Mol Life Sci, 57(10):1423–39, 2000.
- [83] C. D. Rillahan and J. C. Paulson. Glycan microarrays for decoding the glycome. Annu Rev Biochem, 80:797–823, 2011. 19
- [84] L. N. Robinson, C. Artpradit, R. Raman, Z. H. Shriver, M. Ruchirawat, and R. Sasisekharan. Harnessing glycomics technologies: integrating structure with function for glycan characterization. Electrophoresis, 33(5):797–814, 2012. 4
- [85] G. M. Ruiz-Palacios, L. E. Cervantes, P. Ramos, B. Chavez-Munguia, and D. S. Newburg. *Campylobacter jejuni* binds intestinal h(o) antigen (fuc alpha 1, 2gal beta 1, 4glcnac), and fucosyloligosaccharides of human milk inhibit its binding and infection. Journal of Biological Chemistry, 278(16):14112–14120, 2003. 56, 57
- [86] S. S. Sahoo, C. Thomas, A. Sheth, C. Henson, and W. S. York. Glyde-an expressive xml standard for the representation of glycan structure. Carbohydr Res, 340(18):2802–7, 2005.
- [87] J. F. Sanchez, J. Lescar, V. Chazalet, A. Audfray, J. Gagnon, R. Alvarez, C. Breton, A. Imberty, and E. P. Mitchell. Biochemical and structural analysis of helix pomatia agglutinin. a hexameric lectin with a novel fold. J Biol Chem, 281(29):20171–80, 2006. 32
- [88] A. Sarnesto, T. Kohlin, O. Hindsgaul, J. Thurin, and M. Blaszczyk-Thurin. Purification of the secretor-type beta-galactoside alpha 1—2-fucosyltransferase from human serum. J Biol Chem, 267(4):2737–44, 1992.
- [89] K. Sasaki, K. Kurata-Miura, M. Ujita, K. Angata, S. Nakagawa, S. Sekine, T. Nishi, and M. Fukuda. Expression cloning of cDNA encoding a human beta-1,3-n-acetylglucosaminyltransferase that is essential for poly-n-acetyllactosamine synthesis. Proc Natl Acad Sci U S A, 94(26):14294–9, 1997.

- [90] N. Shibuya, I. J. Goldstein, W. F. Broekaert, M. Nsimba-Lubaki, B. Peeters, and W. J. Peumans. The elderberry (*sambucus nigra* l.) bark lectin recognizes the neu5ac(alpha 2-6)gal/galnac sequence. The Journal of biological chemistry, 262(4):1596–601, 1987. 49
- [91] D. F. Smith and R. D. Cummings. Application of microarrays for deciphering the structure and function of the human glycome. Mol Cell Proteomics, 12(4):902–12, 2013. 2, 17, 55, 56
- [92] D. F. Smith, X. Song, and R. D. Cummings. Use of glycan microarrays to explore specificity of glycan-binding proteins. Methods Enzymol, 480:417–44, 2010. 21, 23, 27, 30, 31, 49, 54, 77
- [93] X. Song, Y. Lasanajak, B. Xia, J. Heimbürg-Molinario, J. M. Rhea, H. Ju, C. Zhao, R. J. Molinaro, R. D. Cummings, and D. F. Smith. Shotgun glycomics: a microarray strategy for functional glycomics. Nat Methods, 8(1):85–90, 2011. 56
- [94] X. Song, B. Xia, Y. Lasanajak, D. F. Smith, and R. D. Cummings. Quantifiable fluorescent glycan microarrays. Glycoconj J, 25(1):15–25, 2008. 19
- [95] X. Song, B. Xia, S. R. Stowell, Y. Lasanajak, D. F. Smith, and R. D. Cummings. Novel fluorescent glycan microarray strategy reveals ligands for galectins. Chem Biol, 16(1):36–47, 2009. 19
- [96] X. Song, H. Yu, X. Chen, Y. Lasanajak, M. M. Tappert, G. M. Air, V. K. Tiwari, H. Cao, H. A. Chokhawala, H. Zheng, R. D. Cummings, and D. F. Smith. A sialylated glycan microarray reveals novel interactions of modified sialic acids with proteins and viruses. J Biol Chem, 286(36):31610–22, 2011. 50
- [97] P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, Jr. Stoeckert, C. J., and A. Brazma. Design and implementation of microarray gene expression markup language (mage-ml). Genome Biol, 3(9):RESEARCH0046, 2002. 82
- [98] G. F. Springer and P. R. Desai. Common precursors of human blood group mn specificities. Biochem Biophys Res Commun, 61(2):470–5, 1974. 35
- [99] S. R. Stowell, C. M. Arthur, M. Dias-Baruffi, L. C. Rodrigues, J. P. Gourdine, J. Heimbürg-Molinario, T. Ju, R. J. Molinaro, C. Rivera-Marrero, B. Xia, D. F. Smith, and R. D. Cummings. Innate immune lectins kill bacteria expressing blood group antigen. Nat Med, 16(3):295–301, 2010. 19

- [100] S. R. Stowell, C. M. Arthur, K. A. Slanina, J. R. Horton, D. F. Smith, and R. D. Cummings. Dimeric galectin-8 induces phosphatidylserine exposure in leukocytes through polylectosamine recognition by the c-terminal domain. J Biol Chem, 283(29):20547–59, 2008. 21, 44
- [101] S. Sughii, E. A. Kabat, and H. H. Baer. Further immunochemical studies on the combining sites of lotus tetragonolobus and ulex europaeus i and ii lectins. Carbohydrate research, 99(1):99–101, 1982. 40, 42, 44
- [102] H. Tatenno, A. Mori, N. Uchiyama, R. Yabe, J. Iwaki, T. Shikanai, T. Angata, H. Narimatsu, and J. Hirabayashi. Glycoconjugate microarray based on an evanescent-field fluorescence-assisted detection principle for investigation of glycan-binding proteins. Glycobiology, 18(10):789–98, 2008. 19
- [103] G. Uhlenbruck, G. I. Pardoe, and G. W. Bird. On the specificity of lectins with a broad agglutination spectrum. ii. studies on the nature of the t-antigen and the specific receptors for the lectin of arachis hypogoea (ground-nut). Zeitschrift fur Immunitatsforschung, Allergie und klinische Immunologie, 138(5):423–33, 1969. 35
- [104] G. Uhlenbruck and O. Prokop. An agglutinin from helix pomatia, which reacts with terminal n-acetyl-d-galactosamine. Vox sanguinis, 11(4):519–20, 1966. 32
- [105] T. Urashima, S. Asakuma, F. Leo, K. Fukuda, M. Messer, and O. T. Oftedal. The predominance of type i oligosaccharides is a feature specific to human breast milk. Adv Nutr, 3(3):473S–82S, 2012.
- [106] A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart, and M. E. Etzler. Essentials of Glycobiology. Cold Spring Harbor (NY), 2nd edition, 2009. 1, 3
- [107] A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, J. D. Marth, C. R. Bertozzi, G. W. Hart, and M. E. Etzler. Symbol nomenclature for glycan representation. Proteomics, 9(24):5398–9, 2009. 7
- [108] A. Varki and J. B. Lowe. Biological Roles of Glycans. Cold Spring Harbor (NY), 2nd edition, 2009. 54
- [109] A. Varki and N. Sharon. Historical Background and Overview. Cold Spring Harbor (NY), 2nd edition, 2009. 1
- [110] Claus-W von der Lieth, Thomas Lütteke, and Martin Frank. The role of informatics in glycobiology research with special emphasis on automatic interpretation of ms spectra. Biochimica et Biophysica Acta (BBA)-General Subjects, 1760(4):568–577, 2006. 55

- [111] B. W. Weston, P. L. Smith, R. J. Kelly, and J. B. Lowe. Molecular cloning of a fourth member of a human alpha (1,3)fucosyltransferase gene family. multiple homologous sequences that determine expression of the lewis x, sialyl lewis x, and difucosyl sialyl lewis x epitopes. J Biol Chem, 267(34):24575–84, 1992.
- [112] W. G. Willats, S. E. Rasmussen, T. Kristensen, J. D. Mikkelsen, and J. P. Knox. Sugar-coated microarrays: a novel slide surface for the high-throughput analysis of glycans. Proteomics, 2(12):1666–71, 2002. 19
- [113] A.M. Wu and S. Sugii. Coding and classification of d-galactose, n-acetyl-d-galactosamine, and b-d-gal[i-3(4)]-b-d-glcnac, specificities of applied lectins. Carbohydrate Res., 213:17, 1991. 32, 50
- [114] S. Wu, N. Tao, J. B. German, R. Grimm, and C. B. Lebrilla. Development of an annotated library of neutral human milk oligosaccharides. J Proteome Res, 9(8):4138–51, 2010.
- [115] P. Xuan, Y. Zhang, T. R. Tzeng, X. F. Wan, and F. Luo. A quantitative structure-activity relationship (qsar) study on glycan array data to determine the specificities of glycan-binding proteins. Glycobiology, 22(4):552–60, 2012. 15, 74
- [116] W. S. York, S. Agravat, K. F. Aoki-Kinoshita, R. McBride, M. P. Campbell, C. E. Costello, A. Dell, T. Feizi, S. M. Haslam, N. Karlsson, K. H. Khoo, D. Kolarich, Y. Liu, M. Novotny, N. H. Packer, J. C. Paulson, E. Rapp, R. Ranzinger, P. M. Rudd, D. F. Smith, W. B. Struwe, M. Tiemeyer, L. Wells, J. Zaia, and C. Kettner. Mirage: the minimum information required for a glycomics experiment. Glycobiology, 24(5):402–6, 2014. 82
- [117] Y. Yu, Y. Lasanajak, X. Song, L. Hu, S. Ramani, M. L. Mickum, D. J. Ashline, B. V. Prasad, M. K. Estes, V. N. Reinhold, R. D. Cummings, and D. F. Smith. Human milk contains novel glycans that are potential decoy receptors for neonatal rotaviruses. Mol Cell Proteomics, 13(11):2944–60, 2014. 56, 57, 58, 65
- [118] Y. Yu, S. Mishra, X. Song, Y. Lasanajak, K. C. Bradley, M. M. Tappert, G. M. Air, D. A. Steinhauer, S. Halder, S. Cotmore, P. Tattersall, M. Agbandje-McKenna, R. D. Cummings, and D. F. Smith. Functional glycomic analysis of human milk glycans reveals the presence of virus receptors and embryonic stem cell biomarkers. J Biol Chem, 287(53):44784–99, 2012. 17, 55, 57
- [119] M. X. Zhang, J. Nakayama, E. Hidaka, S. Kubota, J. Yan, H. Ota, and M. Fukuda. Immunohistochemical demonstration of alpha1,4-n-acetylglucosaminyltransferase that forms glcnacalpha1,4galbeta residues in human gastrointestinal mucosa. The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society, 49(5):587–96, 2001. 32, 50

- [120] Z. L. Zhi, A. K. Powell, and J. E. Turnbull. Fabrication of carbohydrate microarrays on gold surfaces: direct attachment of nonderivatized oligosaccharides to hydrazide-modified self-assembled monolayers. Anal Chem, 78(14):4786–93, 2006. 19