The Challenge and Promise of Glycomics

Richard D. Cummings1,2* and J. Michael Pierce1,2*
1Department of Biochemistry, Emory Glycomics Center, Emory University School of Medicine, 1510 Clifton Road NE, Atlanta, GA 30322, USA
2Complex Carbohydrate Research Center, Department of Biochemistry and Molecular Biology, University of Georgia, 315 Riverbend Road, Athens, GA 30602, USA
*Correspondence: rdcummi@emory.edu (R.D.C.), hawkeye@uga.edu (J.M.P.)
http://dx.doi.org/10.1016/j.chembiol.2013.12.010

Glycomics is a broad and emerging scientific discipline focused on defining the structures and functional roles of glycans in biological systems. The staggering complexity of the glycome, minimally defined as the repertoire of glycans expressed in a cell or organism, has resulted in many challenges that must be overcome; these are being addressed by new advances in mass spectrometry as well as by the expansion of genetic and cell biology studies. Conversely, identifying the specific glycan recognition determinants of glycan-binding proteins by employing the new technology of glycan microarrays is providing insights into how glycans function in recognition and signaling within an organism and with microbes and pathogens. The promises of a more complete knowledge of glycomics are immense in that glycan modifications of intracellular and extracellular proteins have critical functions in almost all biological pathways.

Glycoconjugates (GSLs) exert their biological functions through complex molecular mechanisms involving both direct glycan recognition and indirect glycan contributions to conformation and expression of the GSL. Glycans are directly recognized by glycan-binding proteins (GBPcs) (Figure 1). Such interactions of glycans with GPCs can promote cell adhesion, cell-matrix interactions, cellular signaling, glycoprotein folding, and intracellular/extracellular targeting to organelles. In addition, glycans attached to macromolecules exert control through indirect mechanisms on glycoprotein conformation, stability, oligomerization, cell surface resident time, and turnover. Most secreted and membrane proteins are enzymatically glycosylated on one or more amino acids (Apweiler et al., 1999; Steentoft et al., 2013; Van den Steen et al., 1998; Zielinska et al., 2010, 2012), and virtually all nuclear and DNA binding proteins, cytoplasmic enzymes involved in metabolic regulation, and some mitochondrial proteins have the O-β-N-acetylglucosamine (O-GlcNAc) modification on Ser/Thr (Bond and Hanover, 2013; Copeland et al., 2013; Hart et al., 2011; Palaniappan et al., 2013; Yi et al., 2012). An emerging paradigm of modern glycomics recognizes that higher animals express a vast repertoire of glycan structures, which comprise their overall glycomes, and recent studies are unexpectedly revealing that glycans contribute in both general and specific ways to almost all biological regulatory pathways. Thus, as might be predicted, abnormalities in GSL synthesis or turnover are associated with hundreds of different human diseases and disorders, including the congenital disorders of glycosylation (CDGs) and dysglycogenolipopathies, such as congenital muscular dystrophy (Dennis et al., 2009; Filocamo and Morrone, 2011; Freeze, 2013; Hennet, 2012; Ju et al., 2013; Ohtsubo and Marth, 2006). Developing insights into glycan functions and the complexities of glycan structures and conformations represent both the challenge and promise of glycomics, the field of science now recognized as focused on glycans, just as genomics and proteomics are focused on nucleic acids and proteins, respectively.

This article will focus on some of the major challenges and promises of the emerging field of glycomics, both structural and functional. Although the term glycomics usually denotes the chemical aspects of glycoanalysis, we use this term here as a shorthand to denote a broad set of research and knowledge in the chemistry and biology of glycans in terms of structure, function, biosynthesis, role in biology, and disease, etc. Our major emphasis will be on human and animal systems, but we make a point to acknowledge that glycomics and glycoscience represent broad areas of knowledge and research encompassing human and animal biology as well as plant, fungal, and microbial systems.

**Glycan and Glycan Binding Protein Diversity**

Glycans occur as both simple and complex structures in thousands of GSLs, which include glycoproteins, proteoglycans, glycolipids, and free or unconjugated glycans. The factors regulating glycan expression and their molecular and functional roles have long been a supremely challenging puzzle. Moreover, our knowledge of the types of glycans and the number of glycan-amino acid linkages is growing, and the nature of the “core structures” of glycans in glycoproteins, glycolipids, and glycosaminoglycans (GAGs) and glycosylphosphatidylinositol (GPI)-anchored glycoproteins is expanding at an astonishing rate, fueled by development of genomics, proteomics, and mass spectrometry (MS)-based tools (Figure 1). Mammalian glycosomes are built from nine common sugars (Glc, GlicNac, Gal, GalNac, Man, Fuc, GlcA, Xyl, and NeuAc [sialic acid]) with a tenth sugar, IdoA, being created within presynthesized GAGs. At least nine amino acids are known to be glycosylated in nature (Asn, Arg, Ser, Thr, Tyr, Trp, Cys, hydroxylysine, and hydroxyproline) (Spiro, 2002; Stepper et al., 2011). The surface membrane of cells may contain over 10 million glycans linked to Asn and Ser/Thr residues alone, with the concentration of terminal sugars such as sialic acid approaching 100 mM (Wang et al., 2013). The spatial and temporal organization and functions of all these mammalian glycans is largely unclear. While proteomics is defining protein-protein and protein-nucleic interactions that are estimated to be in the many hundreds of thousands (Garma et al., 2012; Venkatesan et al., 2009), with >70% of proteins...
encoded in the human genome containing at least one identifiable protein interaction domain (Liu et al., 2012a; Pawson and Warner, 2007), glycomics has not yet advanced enough to estimate the number of protein-glycan interactions, i.e., what can be termed the “protein-glycan interactome”. While there is a growing list of mammalian glycan binding proteins (GBPs) with defined carbohydrate-binding domains and carbohydrate-binding modules (see Gupta et al., 2012; Taylor and Drickamer, 2011; Varki et al., 2009; Vasta and Ahmed, 2009; and http://www.cazy.org), many new interactions (Tateno, 2010), especially those involving GAGs, are being discovered and will require further mechanistic insights to generally define protein motifs governing glycan recognition. Because the historical term lectin refers to mainly soluble multivalent proteins capable of agglutinating cells and lacking enzyme activity, the commonly understood definition of lectin would preclude many membrane-bound multivalent proteins, antibodies that bind glycans, toxins, microbial adhesins, enzymes that bind glycans, engineered glycosidases, and GAG-binding proteins. Thus, the more general term GBPs encompasses all of these types of proteins, even though the word lectin appears in many databases. We denote in Figure 1 “Direct Glycan Recognition,” where a glycan determinant is directly bound by a GBP, and “Indirect Glycan Effects” that range from glycan-glycan interactions (Handa and Hakomori, 2012), protein solubility, protection from proteolysis and immune surveillance to affecting protein conformation and associations of GSLs in the plasma membrane, for example. Given the growing evidence that the human and individual animal glycomes contain thousands of glycan species (Cummings, 2009) as well as evidence for hundreds of different GBPs encoded in...
Zhang, 2010)( Figure 2), it is possible that the number of protein-animal genomes (Hileman et al., 1998; Schulenburg et al., 2008; Review Chemisty & Biology that infect these animals (van den Berg et al., 2012). The results of glycan microarray studies using different plat- forms and different types of glycans are revealing that many of the GBPs in animals interact with glycans derived from the microbe as well as those expressed by the myriad of pathogens that infect these animals (van den Berg et al., 2012).

Studies on glycan structures and functions are complicated by the facts that most glycans in human and animal tissues remain poorly defined structurally, expression is typically cell type-spe- cific and developmentally- and differentiation-dependent, little is known about the factors governing glycosylation on specific pro- teins and at specific sites, and virtually nothing is known about the overall architecture and topology of glycan expression on and within cells. However, recent developments in this field have been spurred, not only by the important biological functions of glycans being revealed through genetic studies, but also through breakthroughs in technologies involving MS as well as glycan microarray analyses that demonstrate the wide distribution and binding specificities of GBPs. Together, the genetic information and GBP information are being combined to promote a functional understanding of glycans, which has been termed “functional glycomics”. Although the current picture presented here underscores the tremendous challenges in the field, the landscape of glycoscience is rapidly changing, and the field is at the threshold of significant breakthroughs that will rever- berate throughout biology. Such discoveries are highlighting the position of glycans as one of the four major classes of life’s macromolecules, on the center stage of modern biomedical and chemical research (Marth, 2008).

Glycan Microarrays and Glycan Determinants

Protein-glycan interactions were historically studied using the hapten inhibition approach in laborious precipitation or hemagglutination-type assays. Such approaches were far from high throughput, lacked sensitivity, and required large quantities of valuable glycans, thus limiting binding studies to monosaccharides and short oligosaccharides or poorly defined complex glycoconjugates isolated from natural sources. This approach led to the prior paradigm that GBPs have relatively low affinity and lack strong specificity. However, within the last decade and beginning with the introduction of ELISA-type assays in the late 1990s, studies have used microarray technologies in which large libraries of complex glycans are immobilized on sur- faces, akin to nucleic acid arrays. These glycan microarrays have helped to revo-
GBPs, including antibodies, display relatively high affinity interactions with complex glycans, often involving 3–7 monosaccharide residues, which includes branched structures or saccharide modifications such as sulfation or phosphorylation. Researchers are moving away from a focus on the minimal determinants that might inhibit a lectin binding to a single glycan, such as millimolar concentrations of monosaccharides or simple glycans, to the concept of the biological and physiological interactions between relevant naturally occurring GBPs and relatively complex glycan structures (Wang et al., 2013). The concept of “glycan determinants” (Cummings, 2009), which are the “minimal” glycan structures that confer the “maximum” glycan binding affinity, may be thought of as akin to antibody epitopes or glycotopes (Cao et al., 1996). This concept helps us understand the contributions of individual glycan features and biosynthetic pathways to glycan recognition. However, such array studies naturally raise questions as to their physiological relevance, and there are concerns of glycan “presentation” on microarrays and the degree to which this presentation is biologically relevant to their presentation on cell surfaces or native GSLs (Park et al., 2013).

It should be self evident that binding or lack of binding of GBPs to glycans on glycan microarrays or other surfaces does not directly provide evidence of functionally significant interactions, but such results can provide hypotheses to be tested regarding particular glycan functions. In many cases, ligands predicted by glycan microarray analyses have been shown to have physiological relevance, whereas, in some cases, the connection is less clear. Examples of the former are results of virus binding, such as influenza viruses and noroviruses, which bind in microarrays and in vivo to sialylated (Blixt et al., 2004; Stevens et al., 2006; Walther et al., 2013) and blood group H-type glycans (Lindesmith et al., 2003; Shang et al., 2013), respectively. Another example of...
concordance of glycan microarray data and physiological functions are studies on CD22 (Siglec-2), which was shown to bind glycans with NeuAcα2-6-linked sialic acid on cells in vitro (Macauley et al., 2013; Powell et al., 1993; Sgroi et al., 1993), in solution studies (Powell and Variki, 1994), in glycan microarrays (Campanero-Rhodes et al., 2006; Tateno et al., 2008), and in mice (Hennet et al., 1999). A possible example of discordance between glycan microarray data and physiological ligands involves recent studies on Siglec-F in mice. Both human Siglec-8 and its murine homolog Siglec-F preferentially recognize 6'-sulfo-sialyl Lewis X (6'su-SLex - NeuAcα2-3[6'SO3]-Galβ1-4[Fucα1-3][GlcNAc-R]), evidence in part based on binding to immobilized arrayed glycans (Bochner et al., 2005; Kiwamoto et al., 2013; Tateno et al., 2005). This Siglec is expressed in eosinophils, and its binding to 6'su-SLeα has been proposed to be important in mitigating allergic eosinophilic airway inflammation. Studies in vivo indicate, however, that although St3Galα3 mutants lacking the sialyltransferase had diminished Siglec-F binding and more intense allergic eosinophilic airway inflammation (Kiwamoto et al., 2014), deletion of the sulfotransferases capable of generating 6-sulfated galactose, as determined by glycomic analysis, did not appreciably affect Siglec-F binding in a mouse model (Patnode et al., 2013). Thus, in all cases where it is experimentally feasible, it is critical to link the results of glycan microarray analyses for GBP glycan determinants with physiological evidence for such glycan determinants being functionally important. It is also important to note that current glycan microarrays lack the full presentation of structures found in the human and animal glycomes. Thus, many potentially important glycans and glycan determinants are lacking, and lack of binding of a protein to a glycan microarray may simply indicate that the relevant glycan ligand is missing.

Although the number of structurally different glycans in the human glycome is presently unknown, it is likely to be many tens of thousands. The repertoire of glycan determinants in the human glycome, however, is estimable. An approach to this problem is to conceptually assemble partial determinants or segments of glycan structures with each other in biochemically allowed and defined ways based on known biosynthetic pathways and glycan structures (Cummings, 2009; Rademacher and Paulson, 2012; Werz et al., 2007). Such a combinatorial approach permits calculations as to how many such determinants might exist in N- and O-glycans and glycolipids, as well as GAGs. An illustration using partial determinants shows that that there are possibly many thousands of glycan determinants (Figure 4). Few of these glycan determinants are currently available on available microarrays (the current CFG glycan microarray has 611 different glycans); thus, there is a tremendous need for chemical and enzymatic syntheses of thousands of compounds. Fortunately, the NIH has funded the development of small libraries of synthetic glycans as a first step in expanding the availability of glycans to the research community (http://sbir.cancer.gov/funding/contracts/fy2013_09.asp). Coupled with advances in asymmetrical synthesis of glycans (Wang et al., 2013), these efforts should be encouraged, because a large number of glycans are obviously needed for structural and functional studies and as standards for nuclear magnetic resonance (NMR) and MS analyses. In any case, the need for an increased number of synthetic glycans in sufficient quantities to impact glycan analysis and functional studies far exceeds their availability.

Milk and Glycolipid “Metaglycomes” as a Paradigm for Glycomics

Although the human glycome certainly represents a challenge in terms of defining all its structures, the metaglycomes of cells and tissues may be amenable to total analysis. We use the term “metaglycome” to denote a constituent glycome of a specific cell or tissue as well as a type of glycoconjugates summing up all of the human metaglycomes would then define the human glycome. In fact, targeting specific metaglycomes and providing their repertoire of glycans would likely constitute a milestone toward the direction of eventually defining the human glycome. Two such metaglycomes that would be amenable to such analyses are the soluble human milk oligosaccharides or glycans and human glycosphingolipids GSLs. The sizes of both of these metaglycomes have already been estimated, with the HMO comprising several hundred glycans (Bode, 2012; Urashima et al., 2011), and the human GSL metaglycome comprising at least 500 different glycosylated neutral and acid species, based on the LIPID MAPS database (http://www.lipidmaps.org/), without consideration of the aglycone lipid constituents. Importantly, technologies appear to be in place to allow quantitative and qualitative descriptions of both the free glycans in human milk (Bao et al., 2013; Ninonuevo et al., 2006) and releasable glycans from GSL glycomes (Fujitani et al., 2011). Interestingly, both of these metaglycomes have terminal Glc as the nonreducing sugar (lactose-type glycans in milk and lactosylceramide-type glycans in GSLs), so technologies for defining the glycans may be overlapping and synergistic. Recent technological advances indicate that GSL metaglycomes may be approached using intact lipids, which maintains information regarding aglycone dynamics as well (Boccuto et al., 2014).

A common conundrum for the glycomics field is: how will researchers know that they have fully defined a metaglycome? The evidence that a metaglycome is largely, if not fully, defined would be based on the pace of discovery of new glycan species within such metaglycomes in the future; thus, at some point in time, the pace of discovery would lessen to the point that one could estimate that > 95% of the glycans below a certain reasonable size limit, e.g., 5,000 Da, had been described for that metaglycome. The definitive description of a metaglycome to that level, with calculated and statistical reasoning, would go far in assisting the development of technologies and bioinformatic approaches that will provide researchers with their first breakthrough paradigm in defining at least two specific component metaglycomes of the human glycome. A clear challenge to the field of glycomics is that if we cannot fully define the HMO and GSL metaglycomes, how can we possibly imagine defining the glycoprotein-derived metaglycomes, which are clearly much more complex?

The complexity of glycans parallels the technical difficulties in analyses as well as the number of glycan determinants that might exist (Figure 5). Thus, one could imagine that GAGs, with their multitude of potential disaccharide repeating units, would represent the most complex set of glycans in the human glycome, whereas GPI-anchors and human milk oligosaccharides
may represent the relatively least complex set. Thus, future studies in glycomics should focus to some degree on the meta-glycome that are most definable, and functional glycomics can then define the recognition of those glycans by GBPs or the specific roles of those glycans in human physiology and disease. Such a strategy could lead to the identification of the protein-glycan interactome for that particular metaglycome.

In the absence of either knowing all the structures in the human glycome or having all glycan determinants available as synthetic compounds, it is possible to use a “shotgun glycomics” approach, prior to definitive structural characterization, to obtain all the cellular glycans and use fractionated glycan species for functional studies on glycan microarrays and other surfaces. Such a shotgun approach has been successful in many cases for preparing total GSL-derived glycan libraries and other types of natural glycans (Liu et al., 2012b; Song et al., 2011; van Diepen et al., 2012). The combined approach of general glycomics, the release of all the glycans from endogenous GSLs and subsequent structural analysis, with functional glycomics, where individual purified glycans species are probed for their functional recognition by GBPs (Smith and Cummings, 2013), is highly likely to make significant strides to unravel both the mysteries of the glycan structures that make up the glycome and their recognition by endogenous and pathogen-derived GBPs.

Figure 4. Illustration of the Complexity of Glycan Determinants

GBPs often recognize 2–6 linear or branched monosaccharides with additional modifications, such as phosphorylation, sulfation, O-acetylation, etc. Such recognition can be termed a glycan determinant, which is the minimal glycan structure that confers maximum binding. These glycan determinants can be thought of as being assembled from partial determinants, which, in the example shown, are modified galactose, modified GlcNAc, and modified GlcNAc-Man. With the partial determinants shown, which are 6, 9, and 3 in number, respectively, it is possible to assemble these into 162 different glycan determinants. Also shown are two of these as Determinant 1 and 2, which are differently recognized by a GBP or GRM. If one considers all such known partial determinants in human glycans defined to date, it is possible to predict that there are over 5,000 glycan determinants; if the GAG sequences are also included up to pentasaccharides, then there are an additional 10,000 or so. It is likely that this is an underestimate for the total theoretical glycan determinants, because it is likely that other partial determinants will be identified in the future. In addition, it is possible that, in a single branched or linear glycan, the one set of glycan determinants may attenuate the recognition of the same or a different glycan determinant on the same molecule.

Chemistry & Biology 21, January 16, 2014 ©2014 Elsevier Ltd All rights reserved
The different classes of glycans, from the GPI glycans within GPI-anchored glycoproteins to the GAGs, represent different levels of complexity and number and diversity of glycan determinants. Of course, with the complexity and increased numbers of determinants, the analytical difficulties expand tremendously. Thus, a great challenge of glycemic technologies is to discover methods of preparing and analyzing glycans to overcome the challenges of their complexities.

Glycan Recognition Molecules

Over a hundred different plant lectins and a few from invertebrates have been identified, and many, such as Concanavalin A (Jack bean) and Helix pomatia agglutinin (snail), have been utilized successfully to study glycan expression and function. The use of glycan microarrays has allowed the binding determinants of a large number of lectins to be explored, thereby greatly facilitating their utility in testing hypotheses regarding glycan function. The repertoire of human and mouse glycan structures recognized by the plant lectins commonly available is quite limited, however, and often a single lectin binds to multiple glycan structures carrying minimal glycan determinants. In addition, analysis of the binding specificities of commercial plant lectins by the glycan microarray screening of the CFG has demonstrated remarkable inconsistencies among commercial preparations of the same lectin, including variable binding affinities and even inactive preparations. The field of glycomics is at a point where more reliable and better-defined reagents that recognize a breadth of glycan structures are required. Moreover, if investigators encompassing the breadth of biomedical research had these types of reliable reagents available, a significant increase in our knowledge of glycan function would, no doubt, occur. The time is right to consider developing an extensive library of glycan recognition molecules (GRM) with specificities and determinants defined by binding to expansive glycan microarrays. These reagents, if made available to researchers with diverse health-related interests, would greatly accelerate the pace of glycan function discovery in physiology and disease. These reagents could also provide a direct means to target specific glycans for potential therapeutic and diagnostic applications. There are novel platform technologies in the early stages of development that can generate and select for molecules that recognize and bind specific glycan structures. These include antibodies or antibody-like proteins as well as nonprotein molecules such as nucleic acid aptamers (Li et al., 2008) and synthetic lectins (Ke et al., 2012). Because some of these GRMs are not proteins, here we refer to this class as GRM, instead of GBP. Potential technologies for generation and selection of antibodies with specificity for specific glycan epitopes include the engineering of glycans, glycopeptides, and novel adjuvants to elicit high-affinity IgG (Lakshminarayanan et al., 2012), yeast display of scFv antibodies (Zhao et al., 2011), and lamprey antibodies (Yu et al., 2012). In addition, the elucidation of glycan binding specificities of viruses, bacteria, and both plant and animal lectins (using glycan microarrays) as well as the structural knowledge of these GBP plus many glycan-recognizing enzymes provide a structural basis of glycan binding motifs. Knowledge of these motifs, coupled with detailed structural studies of glycan binding domains by crystallography and NMR spectroscopy, as well as molecular simulations of glycan-protein interactions, have led to an expanding understanding of how to engineer glycan recognition determinants (e.g., Feinberg et al., 2013; Ford et al., 2003; Mercer et al., 2013). This knowledge could be exploited to assist the design and selection of GRM with glycan binding determinants and specificities of interest.

The prospect for the availability of libraries of GRM is somewhat analogous to the availability of antibodies directed against specific phospho-peptides for studies of cell signaling pathways and the development of restriction enzymes for molecular biology. Availability of these sequence-specific reagents has allowed the investigation of signaling pathways and nucleic acids by nonexpert investigators with diverse interests. Likewise, GRM would assist in testing for changes of glycan epitope expression and would also provide new tools for assessing functions by potentially blocking or crosslinking target glycans. Production of GRM will involve much more complex challenges than for peptide or phospho-peptide-specific antibodies or restriction enzyme identification because of the diversity of glycan structures, and their precise binding determinants must be determined using glycan microarrays and other approaches, akin to what was done for the cluster of differentiation in immunology (http://www.sciencegateway.org/resources/prow/). Despite the obvious challenges, the availability of defined GRMs to pursue specific biological questions will likely be transformative.

Glycomic Analyses of Various Types of Glycoconjugates and Their Recognition by GBP

A clear challenge of glycomics is the complexity of analytical methods for individual classes of glycans, as discussed here. Each class requires different methods of extraction or glycan release, often different methods of analysis, or combinations of methods, such as high-performance liquid chromatography (HPLC), MS, NMR, etc. (Figure 6). Nevertheless, although complex, each of these methods has been adapted to provide deep insights into glycan structure for that particular glycan class. An all-encompassing method of glycoprotein analysis at the top-down level would be ideal, and coupled with the bottom-up approaches (Hanisch, 2012; Nicolardi et al., 2013), would help to define the relationships of glycans to their protein and lipid carriers and their relative abundance to each other. Moreover, analytical approaches capable of probing the glycomes of living cells, as through biorthogonal methods
Belardi et al., 2012 or noninvasive magnetic resonance approaches, would provide breakthroughs into the dynamical aspects of glycomes in living tissues.

**Glycosphingolipids**

The majority of the current methodologies for preparing glycans beginning with cell/tissue extraction are depicted in Figure 6. Obviously, the challenges for structural determination for the glycans on these diverse classes of GSLs are formidable and call for more facile technologies with automated annotation. Beginning the discussion with GSLs, sophisticated analytical platforms have been developed for their separation and resolution, ranging from thin-layer chromatography to HPLC (Levery, 2005; Suzuki et al., 2011). In many cases, the glycan moieties are released and analyzed separately from the ceramide aglycone, although the accuracy and speed of current mass spectrometers are facilitating the analysis of intact GSLs. In shotgun glycomics for functional studies, the released glycans after separation can be printed to generate glycan microarrays. For site-specific glycosylation and identification of protein carriers, glycopeptides can be generated by proteolysis and then analyzed directly before or after glycan removal.

Figure 6. General Glycomic Strategies

The general strategies for glycomic analyses typically involve isolating or generating free glycans from glycoproteins/proteoglycans and glycolipids. The obtained mixture of glycans can then be derivatized and directly analyzed by MS or derivatized, separated by HPLC and other approaches, and further analyzed by MS or NMR. In shotgun glycomics for functional studies, the released glycans after separation can be printed to generate glycan microarrays. For site-specific glycosylation and identification of protein carriers, glycopeptides can be generated by proteolysis and then analyzed directly before or after glycan removal.

(Belardi et al., 2012) or noninvasive magnetic resonance approaches, would provide breakthroughs into the dynamical aspects of glycomes in living tissues.

**Glycosphingolipids**

The majority of the current methodologies for preparing glycans beginning with cell/tissue extraction are depicted in Figure 6. Obviously, the challenges for structural determination for the glycans on these diverse classes of GSLs are formidable and call for more facile technologies with automated annotation. Beginning the discussion with GSLs, sophisticated analytical platforms have been developed for their separation and resolution, ranging from thin-layer chromatography to HPLC (Levery, 2005; Suzuki et al., 2011). In many cases, the glycan moieties are released and analyzed separately from the ceramide aglycone, although the accuracy and speed of current mass spectrometers are facilitating the analysis of intact GSLs. In shotgun glycomics for functional studies, the released glycans after separation can be printed to generate glycan microarrays. For site-specific glycosylation and identification of protein carriers, glycopeptides can be generated by proteolysis and then analyzed directly before or after glycan removal.

(Belardi et al., 2012) or noninvasive magnetic resonance approaches, would provide breakthroughs into the dynamical aspects of glycomes in living tissues.

**Glycosphingolipids**

The majority of the current methodologies for preparing glycans beginning with cell/tissue extraction are depicted in Figure 6. Obviously, the challenges for structural determination for the glycans on these diverse classes of GSLs are formidable and call for more facile technologies with automated annotation. Beginning the discussion with GSLs, sophisticated analytical platforms have been developed for their separation and resolution, ranging from thin-layer chromatography to HPLC (Levery, 2005; Suzuki et al., 2011). In many cases, the glycan moieties are released and analyzed separately from the ceramide aglycone, although the accuracy and speed of current mass spectrometers are facilitating the analysis of intact GSLs. In shotgun glycomics for functional studies, the released glycans after separation can be printed to generate glycan microarrays. For site-specific glycosylation and identification of protein carriers, glycopeptides can be generated by proteolysis and then analyzed directly before or after glycan removal.
**N-linked Glycoprotein Glycan Analysis and Recognition by GBP**

Asparagine-linked glycans (N-glycans) have been extensively studied in terms of their structures and precise functions (Morgenen et al., 2012). Most of the endogenous Gbps for humans and mice whose ligand specificities are known bind to N-glycans (see http://www.imperial.ac.uk/research/animallectins/cidtd/lectins.html and http://www.functionalglycomics.org/glycomics/publicdata/primaryscreen.jsp. Because of their larger size and extensive branching, the N-glycans are the most complicated non-GAG structures to analyze, because, for example, a tetraantennary glycan can have heterogeneous branches that are even isobaric. These individual branches are difficult to resolve using any number of separation and MS-based platforms of analysis, but progress is being made (Yu et al., 2013a). The analysis of N-glycans, whereby they are released from glycoproteins or glycopeptides using PNGase F or chemical means such as hydrazinolysis, is depicted in Figure 6. Recombinant PNGase F releases all N-glycans found in humans and mice, leaving an Asp acid rather than Asn at the site of attachment to the protein and producing free glycan with an intact reducing terminus. Several methods have been developed to use this enzyme to incorporate $^{18}$O into the Asp where a glycan was attached (Küster and Mann, 1999). Subsequent proteolysis can then produce peptides, some of which have incorporated $^{18}$O, that allow the identification of the specific peptidic structures that had an N-glycan attached. Large scale analysis of N-glycans in tissues and serum has been performed by MS$^n$ using different techniques to enrich for this class of structures. One study utilized four mouse tissues and serum to identify 6,387 N-glycosylation sites on 2,352 proteins; over 99% of these contained the canonical Asn-X-Ser/Thr motif (where X is not Pro); however, a small fraction of these sites actually utilized the Asn-X-Cys motif (Zielinska et al., 2010).

Identification of released N-glycan structures has been pursued by many investigators and methods (reviewed in Han and Costello, 2013; North et al., 2009). It is useful to distinguish between methods that are facile for the discovery of novel glycans or discriminate between isobaric structures, in contrast to those that are optimal for detecting and quantifying structures in a more routine manner with potential for high-throughput analysis. One of the classic techniques used mainly for quantifying mixtures of N-glycans, after release from its protein or peptide, is tagging of the reducing termini of glycans with a fluorescent compound and separating the various species via HPLC or capillary electrophoresis (Doherty et al., 2012). This methodology has been very successfully scaled to high-throughput and automated analysis, allowing large populations of samples to be analyzed. Although, because it is based mainly on retention time differences determined for standards whose structures have been demonstrated by a variety of means, glycans with unusual or unexpected structures may not be resolved. Glycosidase digestions can often aid in structural assignment. A recent study profiling the N-linked structures detected on the major serum glycoproteins of a large population, coupled with a genome-wide association study of the genomes of those whose sera were analyzed, suggested that the transcription factor HNF1α is involved in regulating fucosylation of N-glycans on serum glycoproteins (Thanabalasingham et al., 2013). Fluorescent tagging of glycans still yields one of the lowest thresholds of glycan detection compared to other methods, including various types of MS.

Analyses of released N-glycans by mass spectrometric methods range from simple MALDI-TOF to those that employ permethylation, followed by purification and MS$^n$ techniques. Permethylation assists in extraction of glycans from peptides and other contaminants, neutralizes negative charges, amplifies differences in masses between somewhat similar species, and is able to decrease variability between various molecular ions to be detected by the mass spectrometer. In addition, fragmentation of permethylated glycans produces scats at previous sites of glycosidic linkage, providing additional structural information. Obviously, because of the many isobaric structures found in mixtures of N-glycans, definitive characterization requires fragmentation and extensive analysis (see minimum information required for a glycomics experiment [MIRAGE] below). Quantitation can be done by the total ion mapping method (Aoki et al., 2007), but other methods, also based on those developed for proteomics, have been developed (North et al., 2009; Orlando, 2010). Attempts are being made to automate some of these types of analyses, which will significantly accelerate throughput and move glycan analysis from a specialty to a more common practice.

N-glycans are essential in the quality control of glycoprotein folding and the intracellular trafficking of glycoproteins (Aebi, 2013; Braakman and Bulleid, 2011). Many of the endogenous GBP discovered thus far in humans and mice bind to N-glycans and function in these pathways. The first description of an animal GBP, (Figure 1 Type I), the “Ashwell-Morell receptor,” also termed the hepatic asialoglycoprotein receptor, and founding member of the C-type lectins, is expressed on hepatocytes and binds to N-glycans on glycoproteins when they have terminal galactose due to exposure to neuraminidase (Hudgins et al., 1974). Recent studies have demonstrated that this receptor shows selectivity to ligands that are exposed on desialylated glycoproteins involved in prothrombosis induced by bacterial infection (Grewal et al., 2008). N-glycans also have prominent functions in Type II, indirect glycan effects in regulating glyco-protein function, regulation of glucose transporter (Óhrtubo et al., 2005), receptor half-life on the cell surface (Partridge et al., 2004), and modulation of the effector functions of the Fc moiety of immunoglobulin G (IgG). Evidence has been presented indicating that intravenous Ig is therapeutically anti-inflammatory (Nimmerjahn and Ravetch, 2008) and that 2-6-linked sialic acid on N-glycans of IgG are immunosuppressive (Anthony et al., 2008a; Kaneko et al., 2006) and may bind to DC-SIGN and murine SIGN-R1 (Anthony et al., 2008b; Schwab et al., 2012) to signal the downstream expression of immunosuppressive cytokines and receptors, but more recent studies question some of these prior interpretations (Yu et al., 2013b).

**O-glycan Analysis and Recognition by GBP**

O-glycans in humans and mice range from the shockingly ubiquitous O-GlcNAc modification of Ser/Thr found in intracellular and nuclear proteins (thereby qualifying them to be considered as glycoproteins) (Copeland et al., 2013), to the large family of proteins that contain Ser/Thr-O-GalNAc that is normally extended by other glycans and glycan modifications. Almost all proteins with a predicted signal sequence are also predicted...
to have one or more Ser/Thr-O-GalNAc modifications (Steentoft et al., 2013). Other O-linked glycan modifications include O-Fuc, O-Glc, and O-Man; in fact, one-third of O-linked protein modifications in rat brain have been shown to be O-Man (Chai et al., 1999). Compared to N-glycans, only a few GBPs have thus far been shown to bind O-glycans, suggesting that there are many more of these GBPs to be discovered. The most prominent of these GBPs is P-selectin, which is expressed on activated endothelia and functions in the first step of the inflammatory response involving cell adhesion (Wilkins et al., 1995). Of particular note is the recent study showing a unique O-Man-linked glycan with an unusual repeating polymer of xylose and glucurionate that serves as the ligand for laminin found in basement membranes (Inamori et al., 2013; Yoshida-Moriguchi et al., 2010). This laminin-glycan interaction is responsible for adhesion between many cell types, including muscle and nerve, epithelium, and basal lamina and is affected in cancer cells as well as in several of the CDGs (Freeze, 2013) that affect muscle and nerve function (e.g., Yang et al., 2013). This result also demonstrates that novel glycan structures are still being discovered and that O-glycans can clearly function as ligands for endogenous GBP. The cadherin family of proteins has also recently been shown to express significant levels of O-Man-containing glycans (Vester-Christensen et al., 2013).

Because of the lack of an equivalent of a pan-specific PNGase for O-glycans, structures attached to Ser/Thr residues are typically released by β-elimination via mild base/borohydride or other basic reagents. After release, their separation and quantitation are accomplished by methods similar to N-glycans. There are examples of proteins that have over 1,000 O-GalNAc glycans, whereas others have only a single O-linked GalNAc modification. The diversity of O-GalNAc glycans in a single mucin is staggering when modifications such as sulfation are taken into account; for example, human Muc5ac from patients with cystic fibrosis contains over 260 distinct glycan structures, determined by release, fluorescent tagging, HPLC separation, and MS analysis (Xia et al., 2005). Glycosylation of mucins is often cell- and tissue-specific and altered in disease states (Larsson et al., 2011). The factors that determine selectivity of particular O-GalNAc modifications are only beginning to be understood (Steentoft et al., 2013). The generation of technologies to generate cultured cells that lack particular glycosyltransferases represents a significant step toward understanding the mechanisms that regulate this diversity of glycan structures. Application of MS methods that select specific molecular ions that, when fragmented, serve as signatures for particular O-glycan structures, usually employing tandem MS, can quantify O-glycans in a mixture released from an isolated glycoprotein, serum or cell-derived glycoproteins. This method of selected reaction monitoring and consecutive reaction monitoring or their variants offers the possibility of automation and relatively high throughput of O-glycan analysis and can be used to identify specific glyco-protein glycoforms (e.g., Sanda et al., 2013; Zhang et al., 2012).

**GAG Analysis and Recognition by GBP**

Analysis of GAGs relies mainly on characterization of disaccharides after enzymatic cleavage and further analysis using tandem MS sequencing (reviewed in Zaia, 2013). Clearly, GBPs bind GAGs; the example of antithrombin III binding to heparin is perhaps the most prominent, although these GAG-binding GBPs are often not classified as lectins, because they lack the signature fold that has been identified for various lectin families. Nonetheless, these GBPs are selective in their binding. Using both literature-based and affinity proteomics approaches, the number of GBPs that interact with heparin/heparan sulfate was estimated to be in the hundreds (Ori et al., 2011). Many hyaluronic acid-binding proteins and receptors are also known, notably CD44 and TLR4 (Day and Prestwich, 2002).

**Glycopenomics**

The highest resolution of the glycome would also allow an assignment of individual glycan structures that are expressed at each site on a particular glycoprotein. This assignment is obviously much easier for glycoproteins that express only a few glycans compared to those that express hundreds. The higher the number of glycosylation sites on a protein, the greater the amount of material required for analysis as well as the greater the complexity and time of analysis. Recently, site-specific glycosylation of α-dystroglycan, a glycoprotein with >20 glycosylation sites, has been characterized (Harrison et al., 2012; Stalnaker et al., 2010). Newly developed ion-trap MS techniques to fragment glycopeptides such that glycans can be fragmented by CID, followed by ECD peptide fragmentation and identification, have increased the possibility of assigning glycans to particular amino acid sites in a mixture of glycopeptides (e.g., Halim et al., 2012). Nonetheless, the goal of assigning glycan structures to specific sites on proteins in a complex mixture is still formidable. A novel strategy to metabolically label glycoprotein glycans and GSLs of cultured cells using 15N-glutamine has been developed. This methodology will allow using MS to study the turnover of individual glycans on particular glycoproteins as well as a detailed comparison of glycans in two populations of cells that differ in some way, such as after differentiation, oncogenic transformation, or exposure to a cytokine (Fang et al., 2010; Orlando et al., 2009).

**Glycogenes and Glycotranscriptomes**

Among the initial steps to understand how the glycomes of human and mouse cells are regulated, the first requirement was to produce a list of glycosylation-related enzymes and proteins. The advent of the CaZY database (http://www.cazy.org), an up-to-date, curated collection of enzymes/proteins from all sources that act on carbohydrates, has allowed visualization of the breadth of proteins that recognize glycans or are involved in their metabolism (Nairn et al., 2008). This database, along with several others, has been used to generate a “parts list” of transcripts encoding proteins known or hypothesized (because of sequence identity or similarity) to bind, metabolize, or be directly involved in complex glycan synthesis, breakdown, or transport (Figure 2). The current list of transcripts involved in glycan recognition suggests a total of around 200–210 transcripts in human and mouse. The number of transcripts for putative GBPs, including GAG-binding proteins and others that do not fit into the lectin rubric, is obviously higher. Identification of the amounts of glycotranscripts expressed at any point in time by a particular cell type and how those transcripts change during differentiation, disease progression, or experimental perturbation reveal important clues to understanding which glycans are expressed and their regulation. The nature of the many agents that regulate glycosylation, coupled with competition between biosynthetic enzymes, however, makes it very difficult to extrapolate from transcriptome
data to predict precise levels of the expression of particular glycans. Nevertheless, a recent study focused on the glycomics of mouse embryonic stem cell differentiation to embryoid bodies (EB) and extra-embryonic endoderm (EEE) in terms of both glyco-transcript changes and glycan changes (Nairn et al., 2012). The results suggested that in the vast majority of the cases studied between ESC and EB and between ESC and EEE differentiation, changes in glycosylation transcript levels were consistent with the observed changes in glycan expression.

**Glycobioinformatics**

An overarching goal of bioinformatics focusing on glycomics, also called glycobioinformatics, is to develop and provide tools and algorithms that facilitate the study and identification of glycans, their regulation, and function. This includes software capable of interpreting analytical data to assign glycan structures, databases storing of glycan structures and metainformation, and algorithms linking these structural databases to protein databases, such as UniProt (http://www.uniprot.org). Although analytical tools have not reached the degree of automation as is common in proteomic analysis, they are becoming more essential in the interpretation and display of experimental glycomics structural data. Two widely used tools are GlycanBuilder and GlycoWorkbench (Damerell et al., 2012). The most widely used databases are listed in the Glycomics Portal (http://glycomics.ccr.columbia.edu/GlycomicsPortal/) and allow not only finding and retrieving information about glycan structures, but also information about the interaction of these structures with other macromolecules. For example, searching the CFG and other databases containing glycan array binding data can assist the identification of GBPs that recognize particular glycan structures or substructures (Kletter et al., 2013; McCarter et al., 2013; Choletti et al., 2012).

Although progress has been made in developing and applying such glycobioinformatic tools, many challenges remain. It appears timely to present several proposals for development of bioinformatics tools in order to accelerate glycomics research. (1) An international, open access/open source registry for glycan structures must be developed. In order to facilitate communication between individuals, databases, and scientific literature, an uncurated database that simply associates specific structures with a unique identifier (i.e., accession number) is required. Such a “glycan namespace” will allow each structure, whether it is confirmed or not, to be unambiguously specified by a single identifier. These identifiers can then be used in the communication between tools and databases and will overcome the diverse, incompatible glycan sequence formats currently in use. In addition, a highly curated database containing vetted glycan structures along with metadata (literature references, species information, attachment of the glycan to other macromolecules such as proteins, etc.) is required to facilitate the annotation of the database and provide conceptual links to related biological entities and concepts. UniCarb (http://unicarb-db.biomedicine.gu.se) is making important initial efforts toward such a curated database. Both of the databases must be freely accessible and readable by both scientists and computer programs. The curated database will be most useful if it conforms to MIRAGE standards, as published in *Molecular and Cellular Proteomics* (Kolarich et al., 2013). In addition storing attachment information of glycan structures to glycoconjugates will provide an entry point that can be used by databases of other research fields (e.g., proteomics). (2) Independent databases for experimental glycomics data must also be developed based on various methodologies used in glycomics, e.g., MS, NMR, LC-MS, quantitative RT-PCR/sequencing. These databases, which will be linked via the registries noted above with other data sets, must not just contain the annotated data but also the raw data and metadata about experimental procedures used to generate and annotate the data. The information in these databases must be accessible for researchers but also presented in machine-readable formats and interfaces that allow interaction of the databases with annotation tools, e.g., via web services. To facilitate this communication, at some point there needs to be standardization and agreement on common formats for glycan analysis data (i.e., MS annotation and glycan array data). (3) Manuscripts published in journals and other publications must conform to MIRAGE standards, and clear data deposition requirements, analogous to the requirement to submit coordinate files to the Protein Data Bank for three-dimensional structure reports, must be established for publication. (4) (Semi-)automatic annotation tools, similar to those in use for proteomics, must be developed for interpreting and annotating tandem MS and MS/MS data for glycan structure assignment. These tools must allow sharing of data between users and assist with deposition of data into common registries and databases.

**To Glycoscience and Beyond**

Glycoscience has made spectacular progress in the past few years with the advent of new technologies for exploring glycan structure and function along with insights into the genetic and molecular aspects of glycan expression and regulation. The field is challenged, however, by the complexity and dynamic nature of the glycome and lack of understanding of how glycoconjugates are expressed topologically and temporally. Understanding the functions of glycans and their higher order contributions typically require physiological studies of organisms and identification of altered pathways of anabolism and catabolism in patients (Figure 7). Thus, combinations of studies in single cell systems may fail to identify or even predict these higher order functions. As for any type of biological system, the degree of understanding can be gauge by its predictive ability, and, on this score, glycoscience is truly challenged. One Holy Grail for glycomics researchers is to be able to predict the glycan structure(s) to be found on any particular amino acid site on any glycoprotein of interest; another is for functional glycomics researchers to predict the molecular interactions of a glycan with GBPs and GRMs. At present, our state of knowledge is far from predictive on any level, and we are limited to identifying potential sites of glycosylation on proteins given their amino acid sequences. Defining glycosylation sites and structures of glycans at specific sites is still technically difficult and limited to laboratories with advanced tools, especially considering all the types of amino acid modifications that can occur (Figure 1). A further challenge for glycomics is to understand the links between expression patterns and levels of glycosyltransferases/glycosidases and their localization, spatially and temporally, to particular glycan structures as well as the consequences of such expressions, and that of GBPs, on biological activities. In spite of these challenges, the future is growing brighter as our knowledge grows about
glycan structures and both their direct and indirect functions through studies performed across many biological disciplines. Technological advances in MS and other sequencing methods, glycan synthesis, bioinformatics, and growing knowledge of biological roles of glycans in development, health, and disease, notably the CDGs, provide great hope for the future of glycobiology. The possibility exists to finally tie the knot linking glycans to nucleic acids, proteins, and lipids in the grand theory of every biological, assigning glycans their place as a pillar among those macromolecules that are essential for life.

ACKNOWLEDGMENTS

We gratefully acknowledge input and discussions from our colleagues, Kelley Moremen, Alison Nairn, René Ranzinger, David Smith, Michael Tiemeyer, Lance Wells, Rob Woods, and William York. We also acknowledge research support from NIH Grants to J.M.P. (U01CA128454 and P41GM103490) and R.D.C. (P41GM103694, R24GM098791, U01CA168930, and R01AI01982).

REFERENCES


